1. Link |2008 | Downloaded | The National Cancer Data Base: A Powerful Initiative to Improve Cancer Care in the United States

2. Link |2011 | Downloaded | Biocurators and Biocuration: surveying the 21st century challenges

3. Link | 2013.03 | Downloaded | Comparison of Cases Captured in the National Cancer Data Base with Those in Population-based Central Cancer Registries

4. Link | 2013.08 | Downloaded | Domain knowledge and data quality perceptions in genome curation work

5. Link |2018 | Downloaded | Future of Evidence Synthesis in Precision Oncology: Between Systematic Reviews and Biocuration

6. Link | 2019.03 | Downloaded | Thirty Years of Counting Cancer Cases: The Collective Effort and Efficacy of the National Cancer Database

7. Link | 2019.02 | Downloaded | Incident Cases Captured in the National Cancer Database Compared with Those in U.S. Population Based Central Cancer Registries in 2012–2014

Ann Surg Oncol (2019) 26:1583–1584
https://doi.org/10.1245/s10434-019-07215-z

Annals of
SURGICAL ONCOLOGY
OFFICIAL JOURNAL OF THE SOCIETY OF SURGICAL ONCOLOGY

EDITORIAL – HEALTH SERVICES RESEARCH AND GLOBAL ONCOLOGY

Thirty Years of Counting Cancer Cases: The Collective Effort and Efficacy of the National Cancer Database

Cary Jo R. Schlick, MD[1] and David J. Bentrem, MD, MS[1,2]

[1]Surgical Outcomes and Quality Improvement Center (SOQIC), Department of Surgery, Feinberg School of Medicine, Northwestern University, Chicago, IL; [2]Surgery Service, Jesse Brown VA Medical Center, Chicago, IL

Our scientific community has mapped the human genome and is currently creating a Human Cell Atlas as a table of contents for every cell type in the human body.[1] Despite these advances, we still do not know with certainty the exact number of new cancer cases in the US each year, or the treatments each patient receives. Cancer surgeons, both independently and collectively through the American College of Surgeon's Commission on Cancer, have been a driving force for large-scale cancer research efforts for over 100 years. Surgeons helped establish the American Association for Cancer Research (AACR), which was founded by four surgeons in conjunction with five pathologists and two biochemists, as well as the American Cancer Society (ACS).[2,3]

Dating back to the 1920s, Dr. Ernest Codman, an iconoclastic surgeon from Massachusetts General Hospital, meticulously recorded his patient's treatments and outcomes. He recognized the importance of following cancer patients as an early quality improvement endeavor, and thus developed the first cancer registry consisting of bone sarcoma patients.[4] Shortly thereafter, individual states, beginning with Connecticut, developed statewide cancer registries to record clinical outcomes.[5]

These efforts escalated in the 1980s, at which time cancer surgeons across the country were advocating to advance cancer research through investment in recording cancer outcomes. Around 1987, the late Dr. Gerald Murphy, the national director for the ACS at the time, suggested the Commission on Cancer create a national assessment of cancer care.[6] Dr. David Winchester, then medical director of Cancer Programs at the American College of Surgeons, obtained approval from the College for a National Cancer Database (NCDB). After presenting to the board of directors of the ACS in New York City, there was rigorous debate as to whether to support such an endeavor. In the end, the board voted to support the NCDB, thereby establishing a vital partnership and allowing this concept to become a reality. In 1989, the College, through the Commission on Cancer with support from the ACS, established the NCDB.

For reasons of posterity, it is fitting that the NCDB was established in Chicago, the same city where, a century earlier, Walter Shewhart at Western Electric Company first utilized the concept of recording and quantifying various aspects of complex work to improve performance. He used schematic control diagrams to 'separate signal from noise', thereby establishing the field of industrial quality control as a means to decrease variation in manufacturing processes.[7] The NCDB is maintained today as a quality resource in cancer care, to evaluate treatments and outcomes of patients with cancer.[8]

Previous conventional thinking has been that the NCDB is a hospital-based registry of self-reported cases by Commission on Cancer member institutions, and may not accurately reflect the general US population. In contrast, Surveillance, Epidemiology, and End Results (SEER) data, which is supported by the National Cancer Institute, is an organized, population-based sample of cancer cases designed to assess disease incidence backward over time, and has therefore been seen as the gold standard for cancer studies.[9] However, as the NCDB sample size has grown for clinical and treatment variables, it has encompassed a larger proportion of the population, and thus the likelihood of bias in well-designed studies has decreased.

First Received: 20 January 2019;
Published Online: 6 March 2019

D. J. Bentrem, MD, MS
e-mail: dbentrem@nm.org

At present, approximately 1500 US hospitals submit their cancer data to the NCDB. The article by Mallin et al. establishes the NCDB as capturing 72% of the United States Cancer Statistics (USCS) sample, which is increased compared with prior reports.[10–12] Congress established the National Program of Cancer Registries in 1992 by enacting the Cancer Registries Amendment Act. This act authorized the Centers for Disease Control and Prevention (CDC) to provide funds to States to maintain cancer registries. The USCS is compiled through these state-based cancer registries in 46 states, excluding New Mexico, Iowa, Connecticut, and Hawaii; the SEER database is used to supplement the data for these four states.

In Illinois, cancer cases are collected through mandated reporting by hospitals. The Illinois State Cancer Registry is certified by the North American Association of Central Cancer Registries and has received the gold-standard award for the past 20 years for over 95% or better case ascertainment. It is probably safe to assume the state registries miss up to 5% of cases, which are likely from lower-resourced hospitals.

In the report by Mallin et al., Arkansas is the state with the lowest case coverage by the NCDB, with 19.7% coverage.[10] In terms of disease site, prostate cancer has the lowest coverage. One factor that could contribute to this finding is that the USCS includes prostate cancer cases from Veterans Affairs (VA) hospitals, while cases from VA hospitals in the NCDB are excluded in the study by Mallin et al. Ninety-seven percent of VA cancers are diagnosed in men, with prostate cancer being the most commonly diagnosed malignancy in this population.[13] The decision to exclude VA hospitals from the NCDB analysis will thereby affect the prostate cancer ascertainment rate the most.

The report by Mallin et al., establishing the growing US cancer case coverage rate in the NCDB demonstrates the vital importance of this effort.[10–12] The NCDB is a model of collaboration and collective will on the part of the American College of Surgeons/Commission on Cancer and the ACS. This arduous work needs to continue to document cancer care, recurrence, and disease-specific survival so that as a nation we may continue to work toward answering questions regarding how we treat cancer and to what extent patients benefit from those treatments.

REFERENCES

1. Regev A, Teichmann S, Lander ES, et al. The human cell Atlas. *Elife*. 2017;6:e27041.
2. American Association for Cancer Research. History of the AACR. https://www.aacr.org/ABOUTUS/PAGES/AACR-NARRATIVE-HISTORY.ASPX. Accessed 14 Jan 2019.
3. To Extend Fight on Cancer: Laymen and Physicians Form a Nation-Wide Organization. New York: The New York Times; 1913.
4. Codman EA. The classic: the registry of bone sarcomas as an example of the end-result idea in hospital organization. *Clin Orthop Relat Res*. 2009;467(11):2766-70.
5. Haenszel W, Gurnen MG. The first fifty years of the Connecticut Tumor Registry: reminiscences and prospects. *Yale J Biol Med*. 1986;59(5):475–84.
6. Holleb AI. Quality assurance in the care of the patient with cancer. *Cancer*. 1990;65(11):2419–20.
7. Schlick CJR, Castle JP, Bentrem DJ. Utilizing big data in cancer care. *Surg Oncol Clin N Am*. 2018;27(4):641–52.
8. Hu QL, Ellis RJ, Ko CY. Databases for surgical health services research: national cancer database. *Surgery*. 2018;15(40):1–2.
9. Hankey BF, Ries LA, Edwards BK. The surveillance, epidemiology, and end results program: a national resource. *Cancer Epidemiol Biomarkers Prev*. 1999;8(12):1117–21.
10. Mallin K, BA, Palis B, et al. Incident cases captured in the National Cancer Database compared with those in the U.S. population based Central Cancer Registries in 2012–2014. *Ann Surg Oncol*. 2018;2018-11-2318.R1.
11. Bilimoria KY, Stewart AK, Winchester DP, et al. The national cancer data base: a powerful initiative to improve cancer care in the United States. *Ann Surg Oncol*. 2008;15(3):683–60.
12. Lerro CC, Robbins AS, Phillips JL, et al. Comparison of cases captured in the national cancer data base with those in population-based central cancer registries. *Ann Surg Oncol*. 2013;20(6):1759–65.
13. Zullig LL, Sims KJ, McNeil R, et al. Cancer incidence among patients of the U.S. Veterans Affairs Health Care System: 2010 update. *Mil Med*. 2017;182(7):e1883–91.

Ann Surg Oncol (2019) 26:1604–1612
https://doi.org/10.1245/s10434-019-07213-1

Annals of
SURGICAL ONCOLOGY
OFFICIAL JOURNAL OF THE SOCIETY OF SURGICAL ONCOLOGY

ORIGINAL ARTICLE – HEALTH SERVICES RESEARCH AND GLOBAL ONCOLOGY

Incident Cases Captured in the National Cancer Database Compared with Those in U.S. Population Based Central Cancer Registries in 2012–2014

Katherine Mallin, PhD[1], Amanda Browner, MS[1], Bryan Palis, MA[1], Greer Gay, PhD[1], Ryan McCabe, PhD[1], Leticia Nogueira, PhD[2], Robin Yabroff, PhD[2], Lawrence Shulman, MD, FACP, FASCO[3], Matthew Facktor, MD, FACS[4], David P. Winchester, MD, FACS[5], and Heidi Nelson, MD, FACS[5]

[1]National Cancer Database, American College of Surgeons, Chicago; [2]American Cancer Society, Atlanta; [3]Abramson Cancer Center, Hospital of the University of Pennsylvania, Philadelphia; [4]Thoracic Surgery, Geisinger Medical Center, Danville; [5]Cancer Programs, American College of Surgeons, Chicago

ABSTRACT

Background. The National Cancer Database (NCDB) is a hospital-based cancer registry that includes diagnostic, staging, treatment, and outcomes data for newly diagnosed cancer patients in the United States. The NCDB data include 31 million records for patients diagnosed between 1985–2015. A Participant User File based on a subset of these data has been available to researchers at facilities accredited by the Commission on Cancer since 2010. This study aimed to compare the number of incident cancer cases in the NCDB with a national population cancer registry.

Methods. Incident cancer cases in the NCDB in 2012–2014 were compared with the number of cancer cases in the United States Cancer Statistics data for the 2012–2014 diagnosis years. Comparisons were made by primary site and other factors.

Results. In 2012–2014, the NCDB captured 72% of the cancer cases in the United States, which was slightly higher than the 67% and 69% reported respectively in two prior assessments. Among the top 10 major cancer sites, the highest coverage (80%) was found for breast cancer, and the lowest was found for melanoma of the skin (52%) and prostate (58%). Colon, bladder, and kidney and renal pelvis cancers had relatively high coverage of 71%, 70% and 78%, respectively, whereas lung and bronchus had slightly lower coverage (65%).

Conclusions. The NCDB coverage of U.S. cancer cases has remained relatively high (72%), but differences remain by cancer site and other factors that should be taken into account by users of the NCDB data.

The National Cancer Database (NCDB) is a hospital-based cancer registry that has been collecting data on cancer patients in the United States since 1989.[1] The NCDB is a joint project of the American College of Surgeons Commission on Cancer (CoC) and the American Cancer Society.[2] Since 1998, hospitals accredited by the CoC have been required to submit all their cases with a new diagnosis to the NCDB. Currently, more than 1500 hospitals submit data to the NCDB, which in 2018 included 37 million records for 31 million unique patients with a new diagnosis in 1985–2015.

The NCDB includes diagnostic, staging, treatment, and outcomes information. Two prior studies assessed case completeness and found that the NCDB captured approximately 69% and 67%, respectively, of all U.S. cancer cases compared with the National Cancer Registry data.[3,4] The first study used projected estimates of cancer incidence in 2005 to compare coverage, whereas the second study used 2004–2006 United States Cancer Statistics (USCS) population-based comparison data. The United States Cancer Statistics (USCS) data include population-based cancer incidence data from the Centers for Disease Control (CDC)'s National Program of Cancer Registries (NPCR) and the National Cancer Institute Surveillance, Epidemiology, and End Results (SEER) Program. These registries

First Received: 9 November 2018;
Published Online: 8 February 2019

K. Mallin, PhD
e-mail: kmallin@facs.org

TABLE 1 Case coverage for National Cancer Data Base (NCDB) by cancer site and sex in 2012–2014

Primary site	USCS count	NCDB count	Case coverage	NCDB male count	USCS male count	Case coverage male	NCDB female count	USCS female count	Case coverage female
All cancer sites combined[a]	4,769,679	3,456,127	72.5	1,631,927	2,394,773	68.1	1,824,200	2,374,906	76.8
Male and female breast	706,521	568,498	80.5						
Female breast	700,254	562,876	80.4				562,876	700,254	80.4
Lung and bronchus	649,944	421,478	64.9	218,406	342,271	63.8	203,072	307,673	66.0
Prostate	540,980	315,183	58.3	315,183	540,980	58.3			
Colon excluding rectum	296,070	210,284	71.0	103,127	147,284	70.0	107,157	148,786	72.0
Melanomas of the skin	220,678	114,572	51.9	66,997	130,195	51.5	47,575	90,483	52.6
Urinary bladder	217,214	151,307	69.7	113,773	164,987	69.0	37,534	52,227	71.9
Non-Hodgkin lymphoma	198,421	144,316	72.7	78,727	108,655	72.5	65,589	89,766	73.1
Kidney and renal pelvis	174,204	135,578	77.8	84,734	109,639	77.3	50,844	64,565	78.7
Male and female breast in situ	171,244	141,523	82.6						
Female breast in situ	170,524	140,756	82.5				140,756	170,524	82.5
Brain cranial nerves benign/borderline[b]	160,498	92,697	57.8						
Corpus and uterus, NOS	154,083	126,941	82.4				126,941	154,083	82.4
Thyroid	143,940	117,377	81.6	29,176	35,768	81.6	88,201	108,172	81.5
Leukemias	141,760	89,942	63.5	51,790	82,437	62.8	38,152	59,323	64.3
Pancreas	136,355	102,212	75.0	52,302	69,937	74.8	49,910	66,418	75.1
Rectum and rectosigmoid junction	120,600	91,499	75.9	53,337	70,621	75.5	38,162	49,979	76.4
Liver and intrahepatic bile duct	91,003	64,837	71.3	45,991	65,366	70.4	18,846	25,637	73.5
Brain/cranial nerves malignant[b]	71,987	61,798	85.8						
Stomach	70,410	53,123	75.5	33,126	43,694	75.8	19,997	26,716	74.9
Myeloma	70,081	48,377	69.0	26,964	39,131	68.9	21,413	30,950	69.2
Brain and other nervous system	66,578	56,975	85.6	31,956	37,033	86.3	25,019	29,545	84.7
Ovary	63,866	53,140	83.2				53,140	63,866	83.2
Esophagus	50,059	37,835	75.6	29,622	39,458	75.1	8213	10,601	77.5
Tongue	39,276	30,714	78.2	21,954	28,210	77.8	8760	11,066	79.2
Larynx	37,488	28,660	76.5	22,369	29,743	75.2	6291	7745	81.2
Cervix	37,151	30,696	82.6				30,696	37,151	82.6
Soft tissue including heart	33,779	27,792	82.3	15,218	18,743	81.2	12,574	15,036	83.6
Testis	25,477	18,471	72.5	18,471	25,477	72.5	–	–	–
Hodgkin lymphoma	25,377	19,742	77.8	10,858	14,079	77.1	8884	11,298	78.6
Small intestine	25,027	19,682	78.6	10,233	13,204	77.5	9449	11,823	79.9
Tonsil	23,928	19,026	79.5	15,569	19,652	79.2	3457	4276	80.8
Anus, anal canal, and anorectum	20,302	16,372	80.6	5853	7413	79.0	10,519	12,889	81.6
Other biliary	18,902	16,125	85.3	8673	10,161	85.4	7452	8741	85.3
Other non-epithelial skin	17,678	10,407	58.9	6283	10,739	58.5	4124	6939	59.4
Gum and other mouth	16,792	12,932	77.0	7171	9346	76.7	5761	7446	77.4
Vulva	15,085	12,147	80.5				12,147	15,085	80.5
Salivary gland	13,394	10,047	75.0	5735	7781	73.7	4312	5613	76.8
Gallbladder	12,050	9263	76.9	2957	3844	76.9	6306	8206	76.8
Mesothelioma	9778	7242	74.1	5385	7420	72.6	1857	2358	78.8
Bones and joints	9292	7572	81.5	4249	5234	81.2	3323	4058	81.9
Eye and orbit	8653	5972	69.0	3268	4703	69.5	2704	3950	68.5
Other endocrine including thymus	7356	5701	77.5	2964	3845	77.1	2737	3511	78.0
Nose, nasal cavity, and middle ear	7069	5962	84.3	3568	4272	83.5	2394	2797	85.6

TABLE 1 continued

Primary site	USCS count	NCDB count	Case coverage	NCDB male count	USCS male count	Case coverage male	NCDB female count	USCS female count	Case coverage female
Other female genital organs	7004	6405	91.5				6405	7004	91.4
Hypopharynx	6910	5588	80.9	4469	5604	79.7	1119	1306	85.7
Ureter	6026	4868	80.8	2991	3733	80.1	1877	2293	81.9
Floor of mouth	5953	4790	80.5	3272	4089	80.0	1518	1864	81.4
Peritoneum, omentum, and mesentery	5913	5522	93.4	360	472	76.3	5162	5441	94.9
Lip	5896	2999	50.9	2142	4270	50.2	857	1626	52.7
Oropharynx	5846	4788	81.9	3655	4478	81.6	1133	1368	82.8
Other digestive organs	5785	3959	68.4	1991	2952	67.4	1968	2833	69.5
Nasopharynx	5420	4070	75.1	2853	3850	74.1	1217	1570	77.5
Uterus, NOS	5077	4672	92.0				4672	5077	92.0
Penis	4135	2845	68.8	2845	4135	68.8	.	.	.
Retroperitoneum	3932	3335	84.8	1653	1972	83.8	1682	1960	85.8
Vagina	3931	3061	77.9				3061	3931	77.9
Other oral cavity and pharynx	3517	1801	51.2	1374	2743	50.1	427	774	55.2
Kaposi sarcoma	3437	2000	58.2	1821	3060	59.5	179	377	47.5
Other urinary organs	3328	2326	69.9	1575	2295	68.6	751	1033	72.7
Trachea, mediastinum, and other respiratory organs	1855	1488	80.2	997	1245	80.1	491	610	80.5
Other male genital organs	1108	679	61.3	679	1108	61.3	–	–	–
Pleura	324	231	71.3	130	189	68.8	101	135	74.8

USCS United States Cancer Statistics, *NOS* not otherwise specified

[a]All cancer sites combined exclude in situ, except for urinary bladder

[b]USCS data not available by sex for brain and cranial nerves

cover 100% of the U.S. population.[5] This study aimed to update the NCDB comparisons using USCS data from 2012 to 2014.

METHODS

Incident cancer cases in the NCDB in 2012–2014 were compared with the number of cancer cases in the USCS data for the 2012–2014 diagnosis years. Comparisons were made by primary site, age, sex, race/ethnicity, and the patient's state of diagnosis. Patients living in the U.S. 50 states and the District of Columbia at the time of diagnosis were included in the analyses. The study excluded NCDB patients with a diagnosis and/or, treatment in Veteran's Administration (VA) hospitals. Although the USCS data includes some patient data from VA hospitals, the exact number of VA patients in the USCS is unknown. A survey administered in December 2014 showed 18 NPCR states reporting that they received no data from VA hospitals and 8 states reporting that some but not all VA hospitals were reporting data to their cancer registry (Reda Wilson, personal communication, 8 January 2018).

The USCS cancer data excludes in situ cases except for urinary bladder, and includes separate primary-site categories for invasive and in situ breast. In addition, a separate brain and cranial nerve data file is available, which includes invasive and benign tumors. The other primary site categories in the USCS data use the SEER site recode (ICD-0-3) groups,[6] which are also used by the NCDB. The metric used is the number of cases in the NCDB divided by the number of cases in the USCS.

The number of CoC hospitals providing data in 2012–2014 also was assessed and compared with the number of acute-care short-term hospitals in the United States in 2014 from the Area Health Resource Files.[7] Short-term hospitals are general or specialized hospitals in which the majority of patients stay fewer than 30 days. Long-term facilities as well as psychiatric and specialty hospitals were excluded. All analyses were performed using SAS software (version 9.4; SAS Statistical Institute, Cary NC, USA).

RESULTS

In 2012–2014, the NCDB captured 72.5% of the cancer cases in the United States (Table 1), which is slightly higher than the 67.4% reported for 2004–2006[4] and the 68.6% reported for 2005.[3] Case coverage was slightly higher in 2014 (73.2%) than in 2012 (71.6%) (Table 2).

TABLE 2 Case counts and coverage for United States Cancer Statistics (USCS) and National Cancer Data Base (NCDB) by diagnosis year, race/ethnicity, and age for all cancer sites in 2012–2014

	USCS count	NCDB count	Coverage
Diagnosis year			
2012	1,577,351	1,129,911	71.6
2013	1,595,842	1,157,770	72.5
2014	1,596,486	1,168,446	73.2
Age group (years)			
0–19	45,395	31,088	68.5
20–44	347,139	276,448	79.6
45–54	614,225	484,549	78.9
55–64	1,160,696	874,311	75.3
65–74	1,323,349	947,927	71.6
75–84	897,754	613,806	68.4
≥ 85	381,121	227,998	59.8
Pediatric, AYA age groups (years)			
0–14	31,035	20,218	65.1
15–29	77,867	61151	78.5
30–39	143,497	113,832	79.3
Race/ethnicity			
White	3,995,821	2,902,931	72.6
Black	519,153	381,341	73.5
American Indian/Alaskan Native	27,122	11,230	41.4
Asian/Pacific Islander	146,572	99,056	67.6
Hispanic	361,851	197,457	54.6

AYA Adolescent young adult

Among the top 10 major cancer sites, the highest case coverage (80%) was found for breast cancer, and the lowest for melanoma of the skin (52%) and prostate (58%). Colon, bladder, and kidney and renal pelvis cancers had relatively high coverage of 71%, 70%, and 78%, respectively, whereas lung and bronchus had slightly lower coverage (65%). For malignant brain and other nervous system tumors, coverage was 86%, but for benign/in situ brain and other nervous system tumors, coverage was only 58%.

Comparing coverage by sex, females had slightly higher coverage (77%) than males (68%). This is partially explained by the high coverage for breast cancer compared with the lower coverage for prostate cancer. For most primary sites, the coverage rates for males and females were similar. Relatively high coverage was also found for most female gynecologic sites, such as ovary (83%), corpus and uterus NOS (82%), and cervix (83%).

TABLE 3 Case coverage by patient state for all cancer sites in 2012–2014

Diagnosis state	NCDB count	USCS count	Case coverage
Alaska	5430	8003	67.8
Alabama	48,800	76,344	63.9
Arkansas	9440	47,857	19.7
Arizona	28,321	87,921	32.2
California	258,680	488,056	53.0
Colorado	52,649	66,526	79.1
Connecticut	57,875	61,734	93.7
District of Columbia	7919	8721	90.8
Delaware	14,650	16,951	86.4
Florida	212,902	332,366	64.1
Georgia	112,960	140,670	80.3
Hawaii	14,708	20,838	70.6
Iowa	38,927	52,126	74.7
Idaho	16,766	22,602	74.2
Illinois	169,125	198,066	85.4
Indiana	84,823	98,288	86.3
Kansas	24,892	44,523	55.9
Kentucky	63,189	79,280	79.7
Louisiana	55,244	73,297	75.4
Massachusetts	83,537	108,979	76.7
Maryland	74,876	88,884	84.2
Maine	21,474	25,632	83.8
Michigan	133,458	159,035	83.9
Minnesota	64,482	84,614	76.2
Missouri	78,319	96,947	80.8
Mississippi	33,550	46,837	71.6
Montana	13,289	16,845	78.9
North Carolina	127,999	153,658	83.3
North Dakota	10,604	10,837	97.8
Nebraska	22,924	28,189	81.3
New Hampshire	16,389	23,578	69.5
New Jersey	124,919	149,442	83.6
New Mexico	15,787	26,782	58.9
Nevada	14,592	35,832	40.7
New York	232,515	327,495	71.0
Ohio	168,202	188,193	89.4
Oklahoma	34,026	57,866	58.8
Oregon	45,988	60,954	75.4
Pennsylvania	184,487	233,568	79.0
Rhode Island	16,653	18,451	90.3
South Carolina	58,495	76,332	76.6
South Dakota	9570	13,320	71.8
Tennessee	85,450	103,796	82.3
Texas	197,230	313,294	63.0
Utah	17,571	29,440	59.7
Virginia	99,104	114,342	86.7
Vermont	7226	10,823	66.8
Washington	82,156	105,314	78.0

TABLE 3 continued

Diagnosis state	NCDB count	USCS count	Case coverage
Wisconsin	73,962	94,260	78.5
West Virginia	26,074	34,245	76.1
Wyoming	3919	7726	50.7

NCDB National Cancer Data Base, *USCS* United States Cancer Statistics

Age group comparisons showed the lowest coverage (60%) for the 85 and older age group, with the highest coverage for those 20–64 years of age (77%) (Table 2). The pediatric age group (0–19 years) had 68% coverage, and the adolescent and young adult (AYA) group (15–39 years) had 79% case coverage.

Race and ethnicity comparisons showed that coverage was similar for whites (73%) and blacks (73%), intermediate for Asian/Pacific Islanders (68%), and lowest for Hispanics (55%) and American Indian/Alaskan Natives (AIAN) (41%).

Differences in coverage by the patient's state at diagnosis are influenced by the number of CoC-accredited hospitals in the state or surrounding states. The state of Wyoming, which has only one CoC-accredited facility, had 51% coverage, whereas North Dakota, which has six CoC-accredited facilities, had the highest coverage (98%) (Table 3). Three states and the District of Columbia had more than 90% coverage, and 14 states had 80% to 89% coverage. The lowest coverage occurred in Nevada (41%) and Arizona (32%).

The states also were grouped by census division. The East North Central and New England divisions had the highest coverage, at 85% and 81%, respectively (Table 4). The lowest coverage was found in the Mountain and Pacific divisions, with coverage at 55% and 60%. Census division case coverage for the top 10 cancers mirrored that for all cancers, with New England and East North Central having the highest coverage and the Mountain and Pacific having the lowest coverage. Prostate cancer is one example of this, with case coverage of 73% in the East North Central division and 74% in the New England division, and with 43% coverage in the Mountain region and 47% in the Pacific region.

After VA hospitals were excluded, there were 1475 CoC-accredited hospitals in 2012–2014 compared with 5927 acute-care hospitals in the United States in that year, representing 25% of acute-care hospitals. The distribution of hospital coverage and case coverage is shown in Fig. 1. In general, the higher the hospital coverage was, the higher the case coverage was also. The states in the Northeast had the highest hospital and case coverage, whereas Arkansas, Arizona, and Nevada had low hospital coverage and low case coverage.

DISCUSSION

Overall coverage of cancer cases in the NCDB has remained relatively stable at 72%, with a slight increase above the 67% found in 2004–2006.[4] Case coverage also increased slightly between 2012 and 2014, as did the number of CoC-accredited facilities (excluding VA facilities), which increased from 1455 to 1475, representing approximately 25% of acute-care facilities. Most primary sites have seen slight increases in coverage since 2004–2006[4] or have stayed relatively the same. The primary sites at which cancer is often diagnosed and/or treated in an outpatient setting, such as prostate and melanoma of the skin, had the lowest case coverage. Coverage was slightly higher for females than for males, which is partially explained by the high coverage for breast cancer and low coverage for prostate cancer. In addition, some USCS states include data from VA hospitals, which are not included in the NCDB data.

The distribution of coverage by states increased for most states compared with 2004–2006.[4] Since 2004–2006, 25 states had increases in coverage of 5% or more, with 7 states having increases of 10% or more including New York, Wisconsin, Michigan, Nebraska, West Virginia, Louisiana, and Wyoming. These may reflect newly accredited facilities in these states or surrounding states since 2006.

The states in the Northeast had the highest coverage of both hospitals and cases, and the West and Southwest had both lower case coverage and lower hospital coverage. Data by U.S. census divisions show similar findings, with the highest coverage in the New England (82%) and East North Central (85%) census divisions and the lowest coverage in the Mountain (55%) and Pacific divisions (60%). Similar coverage patterns by major cancer sites also were found by census division. These patterns may also be due to the higher population density in the East, where distance to a CoC facility may be shorter than for patients in lower-density areas in the West and Southwest. Although 93% of the patients received diagnosis, treatment, or both at a facility in their state of residence when their cancer was diagnosed, there was variation by state of residence. The states with a small number of CoC facilities were more likely to have a higher percentage of patients with treatment in a different state from their state of residence (data not shown).

TABLE 4 Case coverage by patient census division and primary site: all sites, and top 10 primary sites

Census division[a]	Primary site	NCDB count	USCS count	Case coverage
New England	All cancer sites combined	203,154	249,197	81.5
Middle Atlantic	All cancer sites combined	541,921	710,505	76.3
East North Central	All cancer sites combined	629,570	737,842	85.3
West North Central	All cancer sites combined	249,718	330,556	75.5
South Atlantic	All cancer sites combined	734,979	966,169	76.1
East South Central	All cancer sites combined	230,989	306,257	75.4
West South Central	All cancer sites combined	295,940	492,314	60.1
Mountain	All cancer sites combined	162,894	293,674	55.5
Pacific	All cancer sites combined	406,962	683,165	59.6
New England	Colon and rectum	16,711	19,767	84.5
Middle Atlantic	Colon and rectum	46,397	60,291	77.0
East North Central	Colon and rectum	55,343	65,669	84.3
West North Central	Colon and rectum	22,011	29,948	73.5
South Atlantic	Colon and rectum	63,796	82,269	77.5
East South Central	Colon and rectum	21,186	28,914	73.3
West South Central	Colon and rectum	27,910	45,871	60.8
Mountain	Colon and rectum	13,275	24,905	53.3
Pacific	Colon and rectum	35,154	59,036	59.5
New England	Corpus and uterus, NOS	7403	8834	83.8
Middle Atlantic	Corpus and uterus, NOS	21,788	26,003	83.8
East North Central	Corpus and uterus, NOS	24,125	25,722	93.8
West North Central	Corpus and uterus, NOS	9283	10,807	85.9
South Atlantic	Corpus and uterus, NOS	25,401	29,429	86.3
East South Central	Corpus and uterus, NOS	7311	8036	91.0
West South Central	Corpus and uterus, NOS	9866	13,418	73.5
Mountain	Corpus and uterus, NOS	5717	9109	62.8
Pacific	Corpus and uterus, NOS	16,047	22,725	70.6
New England	Female breast[b]	43,484	49,323	88.2
Middle Atlantic	Female breast	112,235	131,929	85.1
East North Central	Female breast	123,108	133,396	92.3
West North Central	Female breast	47,229	57,170	82.6
South Atlantic	Female breast	155,563	177,215	87.8
East South Central	Female breast	41,228	50,849	81.1
West South Central	Female breast	58,899	84,105	70.0
Mountain	Female breast	33,470	54,230	61.7
Pacific	Female breast	88,416	132,561	66.7
New England	Kidney and Renal Pelvis	6970	8385	83.1
Middle Atlantic	Kidney and renal pelvis	20,232	24,232	83.5
East North Central	Kidney and renal pelvis	25,043	27,770	90.2
West North Central	Kidney and renal pelvis	10,001	12,436	80.4
South Atlantic	Kidney and renal pelvis	27,625	33,105	83.4
East South Central	Kidney and renal pelvis	9984	11,970	83.4
West South Central	Kidney and renal pelvis	14,777	21,908	67.5
Mountain	Kidney and renal pelvis	6232	10,614	58.7
Pacific	Kidney and renal pelvis	14,714	23,784	61.9
New England	Lung and bronchus	25,248	34,573	73.0
Middle Atlantic	Lung and bronchus	63,054	90,341	69.8
East North Central	Lung and bronchus	79,967	108,418	73.8

TABLE 4 continued

Census division[a]	Primary site	NCDB count	USCS count	Case coverage
West North Central	Lung and bronchus	30,863	46,559	66.3
South Atlantic	Lung and bronchus	95,881	141,170	67.9
East South Central	Lung and bronchus	33,700	51,538	65.4
West South Central	Lung and bronchus	35,053	68,149	51.4
Mountain	Lung and bronchus	15,958	34,224	46.6
Pacific	Lung and bronchus	41,754	74,972	55.7
New England	Melanomas of the skin	7056	11,565	61.0
Middle Atlantic	Melanomas of the skin	18,667	29,890	62.5
East North Central	Melanomas of the skin	19,935	30,739	64.9
West North Central	Melanomas of the skin	9482	16,638	57.0
South Atlantic	Melanomas of the skin	24,835	49,037	50.6
East South Central	Melanomas of the skin	7110	13,052	54.5
West South Central	Melanomas of the skin	6389	16,538	38.6
Mountain	Melanomas of the skin	6545	16,037	40.8
Pacific	Melanomas of the skin	14,553	37,182	39.1
New England	Non-Hodgkin lymphoma	8631	10,752	80.3
Middle Atlantic	Non-Hodgkin lymphoma	23,907	31,118	76.8
East North Central	Non-Hodgkin lymphoma	26,651	30,797	86.5
West North Central	Non-Hodgkin lymphoma	10,887	14,480	75.2
South Atlantic	Non-Hodgkin lymphoma	28,503	37,035	77.0
East South Central	Non-Hodgkin lymphoma	8735	11,428	76.4
West South Central	Non-Hodgkin lymphoma	12,037	20,388	59.0
Mountain	Non-Hodgkin lymphoma	6934	12,016	57.7
Pacific	Non-Hodgkin lymphoma	18,031	30,407	59.3
New England	Prostate	19,392	26,057	74.4
Middle Atlantic	Prostate	47,059	82,935	56.7
East North Central	Prostate	59,935	82,246	72.9
West North Central	Prostate	24,294	36,597	66.4
South Atlantic	Prostate	67,390	112,282	60.0
East South Central	Prostate	22,625	35,835	63.1
West South Central	Prostate	23,901	54,848	43.6
Mountain	Prostate	14,507	33,695	43.1
Pacific	Prostate	36,080	76,485	47.2
New England	Thyroid	7096	9033	78.6
Middle Atlantic	Thyroid	23,270	26,496	87.8
East North Central	Thyroid	19,595	20,756	94.4
West North Central	Thyroid	7257	8941	81.2
South Atlantic	Thyroid	23,029	25,366	90.8
East South Central	Thyroid	6069	7023	86.4
West South Central	Thyroid	10,171	14,332	71.0
Mountain	Thyroid	7197	11,066	65.0
Pacific	Thyroid	13,693	20,927	65.4
New England	Urinary bladder	11,618	13,551	85.7
Middle Atlantic	Urinary bladder	25,890	35,586	72.8
East North Central	Urinary bladder	29,070	35,749	81.3
West North Central	Urinary bladder	10,975	15,314	71.7
South Atlantic	Urinary bladder	30,972	42,637	72.6
East South Central	Urinary bladder	8870	12,715	69.8
West South Central	Urinary bladder	10,021	18,514	54.1

TABLE 4 continued

Census division[a]	Primary site	NCDB count	USCS count	Case coverage
Mountain	Urinary bladder	6905	13,781	50.1
Pacific	Urinary bladder	16,986	29,367	57.8

NOS not otherwise specified

[a]New England: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont; Middle Atlantic: New Jersey, New York, Pennsylvania; East North Central: Indiana, Illinois, Michigan, Ohio, Wisconsin; West North Central: Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, South Dakota; South Atlantic: Delaware, District of Columbia, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, West Virginia; East South Central: Alabama, Kentucky, Mississippi, Tennessee; West South Central: Arkansas, Louisiana, Oklahoma, Texas; Mountain: Arizona, Colorado, Idaho, New Mexico, Montana, Utah, Nevada, Wyoming; Pacific: Alaska, California, Hawaii, Oregon, Washington

[b]Female breast includes malignant and in situ cases

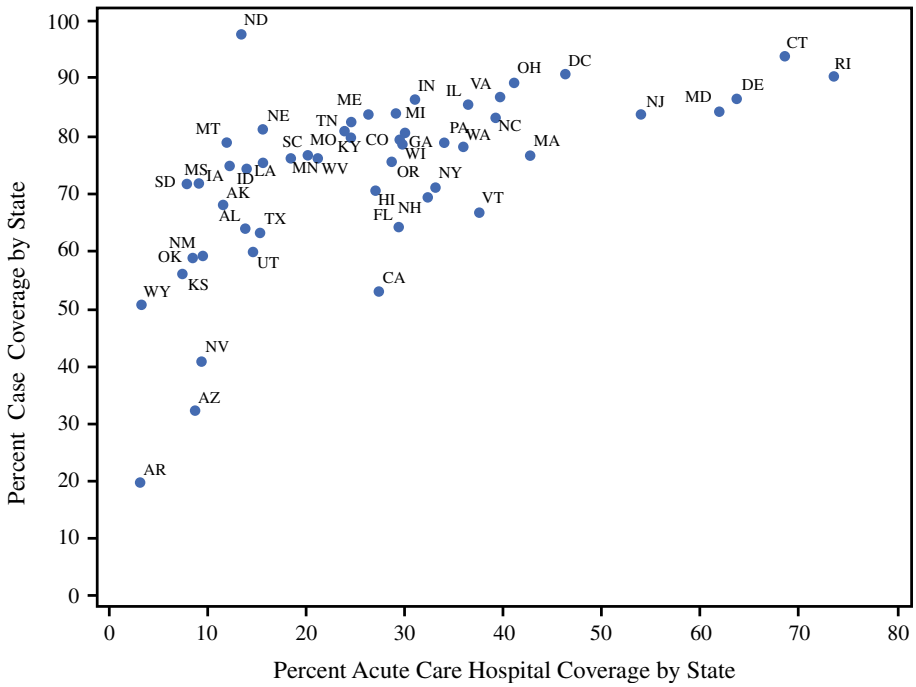


FIG. 1 Percentage of National Cancer Data Base (NCDB) case coverage in 2012–2014 by percentage of NCDB short-term acute-care hospital coverage by state in 2014

Comparisons by race for all cancer sites showed that whites and blacks have a similar coverage rate (73%). Asians and Pacific Islanders had slightly lower coverage (68%) than whites and blacks. The lowest coverage (41%) was found for American Indians and Alaskan natives. In 2010, the states with the highest American Indian populations in the United States were California, Oklahoma, Arizona and Texas,[8] which had relatively low coverage rates in the NCDB. Also, the SEER and NPCR cancer registries are able to link their data to the Indian Health Service Administrative records to help identify the American Indian and Alaskan native population.[9]

Hispanic coverage in the NCDB was only 56%. The lower coverage rates for Hispanics, Asians, and Pacific Islanders may be partially due to the fact that North American Association of Central Cancer Registries (NAACCR) uses an algorithm that identifies Hispanics as well as Asians and Pacific Islanders, which is not available to hospitals.[9] In addition, California, Texas, Arizona, and New Mexico have the highest percentage of Hispanics in the United States,[10] and these states also have lower case coverage in the NCDB.

Lower coverage rates for the oldest group may be partially due to the fact that the state registries also identify a

small percentage of cases based on death certificates alone.[11] Although no more than 5% of cases in the registry are allowed to be based on death certificate alone, these may be more likely to occur for elderly patients. The NCDB data do not include death certificate-alone cases, and this may have a slight impact on the coverage of the oldest cancer patients.

The fact that the NCDB case coverage has remained relatively stable during the past decade is relevant to users of the NCDB Participant User Files (PUFs). These files are a de-identified subset of the NCDB data available to researchers at CoC facilities.[12] However, these users should also be aware of some differences in case coverage by primary site, geography, age, race, and Hispanic origin documented in this report. Users of PUFs should be aware of these differences when analyzing data for any of these subgroups.

ACKNOWLEDGMENT The USCS data were provided by central cancer registries participating in the Centers for Disease Control (CDC)'s National Program of Cancer Registries (NPCR) and/or NCI's Surveillance, Epidemiology, and End Results (SEER) Program and submitted to CDC and NCI in November, 2016. Reda Wilson from the CDC provided useful comments and feedback.

REFERENCES

1. Steele GD, Winchester DP, Menck, HR. The National Cancer Data Base. *Cancer.* 1994;73:499–504.
2. American College of Surgeons. National Cancer Database. Retrieved 11 July 2018 at https://www.facs.org/quality-programs/cancer/ncdb.
3. Bilimoria K, Stewart AK, Winchester DP, Ko, CY. The National Cancer Data Base: a powerful initiative to improve cancer care in the United States. *Ann Surg Oncol.* 2008;15:683–90.
4. Lerro CC, Robbins AS, Phillips JL, Stewart AK. Comparison of cases captured in the National Cancer Data Base with those in population-based Central Cancer Registries. *Ann Surg Oncol.* 2013;20:1759–65.
5. U.S. Cancer Statistics Working Group. *United States Cancer Statistics: 1999–2014 Incidence and Mortality Web-based Report.* U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute, Atlanta, GA, 2017. Retrieved 26 July 2018 at www.cdc.gov/uscs.
6. National Cancer Institute. Surveillance, epidemiology, end results program. Site Recode. Retrieved 30 May 2018 at https://seer.cancer.gov/siterecode/.
7. Health Resources and Services Administration Data Warehouse. Bureau of Health Workforce (BHW). National Center for Health Workforce Analysis (NCHWA). Area Health Resources Files. State and National Level Data. Retrieved 2 May 2018 at https://datawarehouse.hrsa.gov/data/datadownload.aspx.
8. United States Census Bureau. U.S. Department of Commerce. The American Indian and Alaska Native Population 2010. Retrieved 31 May 2018 at https://www.census.gov/prod/cen2010/briefs/c2010br-10.pdf.
9. Centers for Disease Control and Prevention. United States Cancer Statistics USCS. Interpreting Race and Ethnicity in Cancer Data. Retrieved 31 May 2018 at https://www.cdc.gov/cancer/npcr/uscs/technical_notes/interpreting/race.htm.
10. Henry J. Kaiser Family Foundation. State Health Facts. Population distribution by race/ethnicity. Retrieved 30 May 2018 at https://www.kff.org/other/state-indicator/distribution-by-raceethnicity/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D.
11. Centers for Disease Control and Prevention. United States Cancer Statistics. USCS Publication Criteria. Retrieved 31 May 2018 at https://www.cdc.gov/cancer/npcr/uscs/technical_notes/criteria.htm.
12. American College of Surgeons. Quality Programs. National Cancer Database. Participant User Files. Retrieved 1 June 2018 at https://www.facs.org/quality-programs/cancer/ncdb/puf.

Future of Evidence Synthesis in Precision Oncology: Between Systematic Reviews and Biocuration

Simina M. Boca
Orestis A. Panagiotou
Shruti Rao
Peter B. McGarvey
Subha Madhavan

Author affiliations and support information (if applicable) appear at the end of this article.

Corresponding author: Simina M. Boca, PhD, Innovation Center for Biomedical Informatics, Georgetown University Medical Center, 2115 Wisconsin Ave NW, Suite 110, Washington, DC 20007; e-mail: smb310@georgetown.edu.

Precision medicine refers to the tailoring of interventions to patients using approaches that go beyond traditional clinical characteristics (eg, age, sex, disease, symptoms, and medical history) by considering biomarkers consisting of genetic characteristics or molecular profiles.[1] In particular, precision oncology (PO)[2] includes both the development of novel cancer therapies targeting specific changes that occur in some individuals but not others (eg, inherited mutations or somatic mutations that arise in the carcinogenesis process) and the stratification of individuals according to existing interventions (eg, via screening programs tiered by genetic risk or chemotherapies predicted to work for certain molecular profiles), which is expected to show improved outcomes in the resulting subgroups. Herein, we discuss two approaches for synthesizing evidence for or against the use of specific PO interventions, namely, systematic reviews (SRs) and biocuration, and argue that their engagement with each other could facilitate the timely delivery of appropriate PO interventions.

Cancer is known to be highly heterogeneous, and clinical responses differ among patients. PO thus includes a wide variety of applications, including both the use of targeted therapies against particular changes occurring in individual tumors—such imatinib against the *BCR-ABL* fusion in chronic myeloid leukemia,[3] vemurafenib[4] and dabrafenib[5] against specific *BRAF* mutations in melanoma, and trastuzumab against HER2-postive breast cancer[6]—and the assignment between existing therapy classes via biomarkers that are not direct drug targets. This can be based either on the activation of certain pathways, as in the case of EGFR inhibitors in *KRAS*-wild type colon cancer, [7,8] or on gene expression signatures or other molecular characteristics (ie, Onco*type* DX test for breast cancer, used to predict recurrence and likely benefit of chemotherapy).[9,10] These biomarkers are well-established and have been approved by the US Food and Drug Administration for specific cancer types.

The biomarkers discussed in the previous paragraph are classified as having the highest level of evidence for clinical use in society guidelines (eg, those jointly put forward by the Association for Molecular Pathology, ASCO, and the College of American Pathologists).[11] To be meaningful for clinical decision making, biomarkers should have high predictive performance (ie, they can be used to stratify patients into treatment groups with differential outcomes). Most research to date, however, has assessed whether biomarkers are prognostic (ie, are associated with a clinical outcome for untreated or standard-of-care patients).[12] It is often the case that the supporting literature includes few if any randomized control trials.[13] As a result, lower levels of evidence may be assigned, eg, to biomarkers that predict response to a treatment based on well-powered studies evaluated by expert consensus or that predict response in a different tumor type than that being studied.

Many conceptual frameworks have been proposed to facilitate clinical decision making by helping end users (especially patients and clinicians) assess the evidence for or against the use of medical tests, including biomarkers.[14,15] With minor variations, the different frameworks involve a stepwise approach that starts with analytic validity (laboratory reproducibility), proceeds to clinical validity (predictive accuracy, including sensitivity, specificity, and positive and negative predictive values) and clinical utility (use of the biomarker leads directly to improvements in outcomes like survival and quality of life), and eventually assesses the cost effectiveness of the biomarker. Even for biomarkers that successfully reach these milestones, implementation in clinical practice may still be challenging because of additional barriers.[16]

Table 1. Comparison of Key Features of Systematic Reviews and Biocuration

Feature	Systematic Reviews	Biocuration
Overall approach	Top down; starts with a key scientific question and develops evidence base around it	Either top down or bottom up; starts with a number of scientific questions or allows individual articles to be curated and slotted into categories
Evidence assessment	Exhaustive literature searches and systematic assessment of evidence	Multiple frameworks exist
Presence of bias	Systematic assessment of the risk of bias	Higher chance for bias by selecting studies with significant findings
Delivery approach	Peer-reviewed article or whitepaper	Database of entries may include summary of findings and links to original studies and other resources
Research approach	Team usually assembled at the start of the project	Often relies on crowdsourcing and expert review of entries
Updating	Significant lag	Closer to real time, depending on availability of curators
Authors or curators	Multidisciplinary teams, often exclude or reduce role of primary study authors or subject matter experts due to risk of bias	Scientists and clinicians, who are frequently subject matter experts, from a variety of fields, and often with an expert panel to resolve conflicting interpretations

As with all tests in clinical practice, tests related to biomarkers are delivered within given health care settings, and currently, little structured evidence exists in regard to how PO can successfully be implemented for particular sets of patients, with different degrees of access to health care, across diverse health care systems. This requires systematic, timely, and accurate collection and evaluation of the evidence base for the benefits and harms of specific PO interventions. As this evidence accumulates from scientific and clinical studies, two approaches dominate its synthesis and assessment. The first is the traditional evidence appraisal that applies the well-established concepts of evidence-based medicine (EBM).[17,18] EBM integrates a physician's clinical experience with scientific evidence that has undergone an SR, which involves extensive surveying of the literature followed by a synthesis of the primary studies.[19] The second is biocuration, which refers to the distillation and integration of biologic information from scientific literature and large data sets using database- or research field-specific, controlled vocabularies or ontologies; it is a cornerstone of bioinformatics.[20] Both disciplines have the same overarching goal of bringing to patients only the interventions proven to be effective by carefully balancing benefits and harms, doing so in different ways; SRs start with a particular question and systematically develop the evidence base around it, whereas biocuration may be either top-down or bottom-up and usually involves more real-time updating. We compare key aspects of these two approaches in Table 1 and summarize important feature of specific curated databases in Appendix Table A1.

The goal of SRs is to allow a comprehensive and global view of the available evidence base on a particular question of interest by analyzing primary studies that meet prespecified eligibility criteria via explicit, reproducible methodologies that minimize bias. The typical steps involved in an SR include identification of the clinical question, specification of eligibility criteria, systematic search for all studies that meet these criteria, extraction of evidence from eligible studies and assessment of their methodologic rigor, and analysis and qualitative and/or quantitative synthesis (meta-analysis) of the extracted data.[19] An SR is typically structured around the PICOTS elements: the population toward which the findings will be applicable, the intervention, the comparisons being performed, the clinical outcomes, the timing of the eligible studies, and the clinical setting. During the course of decades, a toolbox of methods has been developed in SRs and EBM to evaluate key elements of published studies including systematic biases, statistical precision, applicability to a target clinical setting, and others.[21-23] Evidence for specific interventions on the basis of supporting SRs is generally considered strong. For example, a joint Centers for Disease Control and Prevention/National Cancer Institute framework ranks

clinical practice guidelines on the basis of SRs that support the use of a genomic test as tier one ("ready to implement in clinical practice"[24]) and those that are not based on SRs as tier two ("may be useful in the context of informed clinical decision making"[24]).

In general, SR teams are multidisciplinary, consisting of clinicians, content experts, methodologists, statisticians, and librarians. Because researchers may often be influenced by their own investment in the field when interpreting the evidence base,[25] ideally, parties involved in a SR should have no personal vested interests in the topic of the SR. Although content experts are critical for the clinical interpretation of primary studies and for putting them in context with the broader evidence, their role as authors in SRs has been debated.[26] However, since SRs have various degrees of complexity, relevant sponsors often make decisions on a case-by-case basis by thoroughly reviewing the potential conflicts of individual researchers. For example, the Agency for Healthcare and Research Quality Evidence-Based Practice Centers Program has a set of guidelines for how different types of conflicts should be handled and alleviated.[27] Given that precision medicine is a relatively new and rapidly evolving field with potentially more complex statistical analyses and contextual questions compared to traditional studies, including and relying on content experts is more critical.

Synthesis of evidence from PO studies encounters a number of specific challenges. In some cases, the biomarker of interest may be examined as predictive for different treatments, but the number of studies considered for each treatment may be small. Consider for instance a Cochrane review of first-line treatments in individuals with *EGFR*-mutated noncurable stage IIIB to IV nonsquamous non–small-cell lung cancer, which included 19 RCTs.[28] The overall conclusion was that the tyrosine kinase inhibitor therapies (erlotinib, gefitinib, and afatinib) led to improved progression-free survival but not to improved overall survival, whereas the monoclonal antibody cetuximab did not show any improved outcomes. While this study included 19 RCTs total, comparing targeted therapies to chemotherapy or best supportive care, the number of RCTs conducted for each therapy was between two and eight and the study designs, outcome measures, and analyses performed and reported varied. These factors led to each meta-analysis considering no more than five RCTs. SRs have also identified existing evidence gaps for other prevision medicine interventions, such as the lack of robust evidence for tailoring smoking cessation interventions on the basis of germline genetic variation.[29]

Many issues with the SR of evidence for PO interventions will be solved in time, as more studies accumulate, but it is important to note that the desire for new therapies may be at a historical high given the stated promise of precision medicine and PO in particular. However, another challenge that will linger is that the definition of precision medicine will continue to make it difficult to satisfy the PICOTS framework as, for example, more drugs are tested separately in different subsets of patients (eg, if the same biomarker is considered for different tumor types) or molecular tests rapidly evolve to add more biomarkers or change how they are tested (eg, protein expression, gene expression, or DNA amplification).[13] New RCT designs such as umbrella and basket trials are now being implemented to meet some of these challenges. Although other study designs may provide additional valuable information on specific treatments, the question of clinical utility, in particular, is difficult to answer in the absence of evidence from RCTs.

Biocuration identifies and summarizes biomedical results, including potentially those from SRs, into bioinformatic databases, often by using controlled vocabularies and prespecified standards.[20] A text summary of the evidence related to specific scientific questions may also be included, along with links to the original studies or other resources. In particular, medical curation focuses on disease associations and genetic factors and includes efforts like the OMIM (Online Mendelian Inheritance in Man) database[30] as well as specialized databases for particular genes, variants, and diseases. ClinVar, a public archive of relationships among medically important germline and somatic variants and human phenotypes, was launched in April 2013.[31] Biocuration generally provides more immediate updating than formal SRs, but the lack of a unified systematic framework may lead to only partial or inconsistent results. For example, the ClinVar archive contains clinical interpretations of genetic variants submitted by laboratories and expert

curators, resulting in a wealth of information but also in issues such as contradictory interpretations and many variants of uncertain significance.[32] The ClinGen (Clinical Genome Resource) initiative aims to resolve some of these problems and answer the critical questions of clinical validity, disease causality (pathogenicity), and clinical actionability by standardizing data collection and sharing and implementing an approach for consensus among expert curators.[32] Many other cancer-specific databases now exist for synthesizing evidence for PO approaches,[33-37] leading ClinGen to recently develop recommendations and guidelines for defining cancer somatic variants on the basis of their diagnostic, prognostic, and predictive roles, using evidence of their significance and clinical utility.[38] In a translational field such as PO, where molecular genomics and clinical practice come together, a particular challenge for biocuration is that each content area has its own detailed terminology (eg, specific diseases and syndromes can be described using ICD-10 codes[39] and SNOMED terms[40]). However, mapping these terminologies to specific causal mutations, as well as to biomarkers and tratments, can often be a challenge.

Biocurators come from a wide variety of backgrounds, usually in biology, biochemistry, or medical genetics, and increasingly have interdisciplinary training that includes computer and information sciences, with subject matter experts playing an important role in the field of biocuration.[20,41] A growing number of clinical laboratories in hospitals and community clinics are now also conducting molecular diagnostic testing to identify sequence variants and inform treatment decisions for patients, in which case the curation is often performed by molecular pathologists as well as clinicians. The expert panels used by resources such as ClinGen to resolve conflicting interpretations and curate variants of unknown significance include medical professionals, medical geneticists, clinical laboratory diagnosticians and molecular pathologists that have a long standing scope of work in the disease gene in question.

An example of a curated cancer database is the recently released CIViC (Clinical Interpretation of Variants in Cancer) database,[36] which provides associations between drugs and genes or variants in specific cancer types. Like other such repositories, CIViC relies on crowdsourcing as a way of ensuring biocuration support but only accepts expert-reviewed contributions. It provides a large amount of data, including at the time of this commentary results from 1,077 published studies. However, the authors state the presence of a bias toward positive associations with treatment outcomes (91% of records show support for the use of a therapy), which illustrates the absence of a systematic, unbiased framework for identifying the evidence base. Additionally, it does not include quantitative results, such as differences in overall survival between treatment groups. This approach is thus quite different from the traditional SR approach. Given that CIViC considers many different types of studies and that studies of cancer drugs may have many different designs, it is true that it would be extremely challenging to have a fixed list of terms that would be applicable to all studies and easily translatable into PICOTS terminology.

In general, literature searches for SRs are typically more exhaustive and thorough than for curated databases, where eligible studies tend to be identified by experts in a given field. Although currently this may have little direct impact on identifying relevant studies for biocuration, given the small number of clinical PO publications, this will eventually become a major issue in the next years, because the field will need to deal with increasing amounts of published and unpublished evidence. SRs also tend to have more well-defined a priori eligibility criteria that a study has to meet to be considered in the evidence synthesis; in contrast, biocuration approaches show more variability and rely more on individual experts in the field. Although biocuration currently lacks a standardized and widely accepted framework to assess the quality of published evidence, several efforts are under way to change this.[38,42-44]

Beyond trying to develop standardized frameworks for biocuration, a number of other approaches for improving curated databases have been considered, including the use of specific incentives. For example, databases can have annotation jamborees[45] with invited experts, which may result in publications. Some large projects such as UniProt[46] and ClinVar[31] employ curators and train them in specific guidelines and processes. While mandatory submissions are successfully used for archival databases or repositories such as the Gene Expression Omnibus,[47] curated

databases generally prefer only high-quality representative publications, meaning that they would not encourage this approach. The emphasis on a systematic survey of the literature which includes non-significant results that is present in SRs is also not an explicit part of the biocuration philosophy. We suggest, that moving forward, it would be extremely helpful for biocurators to be made aware of the SR process and vice-versa. For example, if the EBM community engages the biocuration community, they may find ways of incorporating more elements of the PICOTS framework into existing and future databases.

There is often a tension between systematic, unbiased syntheses and assessments versus timely and accessible curated information, and the highly promising field of PO is a clear example that shows the difficulties in finding a good solution that considers both of these dimensions. In particular, SRs often tend to focus on the first aspect and biocuration on the second. We believe that these two approaches to synthesizing evidence must come together so that their members can discuss their specific philosophies, approaches, and products, instead of continuing to proceed on parallel paths. Expertly curated databases can, for instance, serve as one of the inputs into SRs, because they may provide publications initially missed by the reviewers. Conversely, SRs, including those that appear as detailed whitepapers, should always be included in curated databases. Curated databases will continue to be an initial first stop for many researchers and clinicians, especially because SRs may lag years behind current studies, and decisions often need to be made in cases where SRs are lacking. However, more training is needed for users to understand that, for example, a higher risk of bias exists for finding more significant associations when consulting a curated PO database than when using a SR. Both communities have much to learn from each other, and their collaboration could result in substantial improvements in evidence synthesis in PO, in terms of both quality and speed.

DOI: https://doi.org/10.1200/PO.17.00175
Published online on ascopubs.org/journal/po on April 10, 2018.

AUTHOR CONTRIBUTIONS

Conception and design: Simina M. Boca, Subha Madhavan

Collection and assembly of data: Simina M. Boca, Shruti Rao

Data analysis and interpretation: Orestis A. Panagiotou, Peter B. McGarvey

Manuscript writing: All authors

Final approval of manuscript: All authors

Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/po/author-center.

Simina M. Boca
Research Funding: Symphogen (Inst)

Orestis A. Panagiotou
No relationship to disclose

Shruti Rao
Research Funding: Symphogen (Inst)

Peter B. McGarvey
No relationship to disclose

Subha Madhavan
Leadership: Perthera
Stock and Other Ownership Interests: Perthera
Consulting or Advisory Role: Perthera

Affiliations

Simina M. Boca, **Shruti Rao**, **Peter B. McGarvey**, and **Subha Madhavan**, Georgetown University Medical Center, Washington, DC; and **Orestis A. Panagiotou**, Brown University School of Public Health, Providence, RI.

Support

Supported by the National Institutes of Health BD2K Program via Grant No. U01HG008390 and the Ruesch Center for the Cure of Gastrointestinal Cancers.

REFERENCES

1. National Research Council: Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. Washington, DC, National Academies Press, 2011
2. Ford JM: Precision oncology: A new forum for an emerging field. JCO Precis Oncol 1:1-2, 2017
3. O'Brien SG, Guilhot F, Larson RA, et al; IRIS Investigators: Imatinib compared with interferon and low-dose cytarabine for newly diagnosed chronic-phase chronic myeloid leukemia. N Engl J Med 348:994-1004, 2003
4. Chapman PB, Hauschild A, Robert C, et al; BRIM-3 Study Group: Improved survival with vemurafenib in melanoma with BRAF V600E mutation. N Engl J Med 364:2507-2516, 2011
5. Hauschild A, Grob J-J, Demidov LV, et al: Dabrafenib in BRAF-mutated metastatic melanoma: A multicentre, open-label, phase 3 randomised controlled trial. Lancet 380:358-365, 2012
6. Slamon DJ, Leyland-Jones B, Shak S, et al: Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. N Engl J Med 344:783-792, 2001
7. Amado RG, Wolf M, Peeters M, et al: Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. J Clin Oncol 26:1626-1634, 2008
8. De Roock W, Piessevaux H, De Schutter J, et al: KRAS wild-type state predicts survival and is associated to early radiological response in metastatic colorectal cancer treated with cetuximab. Ann Oncol 19:508-515, 2008
9. Marchionni L, Wilson RF, Wolff AC, et al: Systematic review: Gene expression profiling assays in early-stage breast cancer. Ann Intern Med 148:358-369, 2008
10. Calonge N, Klein RD, Berg JS, et al: Recommendations from the EGAPP Working Group: Does the use of Oncotype DX tumor gene expression profiling to guide treatment decisions improve outcomes in patients with breast cancer?. Genet Med 18:770-779, 2016
11. Li MM, Datto M, Duncavage EJ, et al: Standards and guidelines for the interpretation and reporting of sequence variants in cancer: A joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. J Mol Diagn 19:4-23, 2017
12. Simon R: Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. Per Med 7:33-47, 2010
13. Rao S, Beckman RA, Riazi S, et al: Quantification and expert evaluation of evidence for chemopredictive biomarkers to personalize cancer treatment. Oncotarget 8:37923-37934, 2017
14. Lijmer JG, Leeflang M, Bossuyt PM: Proposals for a phased evaluation of medical tests. Med Decis Making 29:E13-E21, 2009
15. Ioannidis JP, Khoury MJ: Improving validation practices in "omics" research. Science 334:1230-1232, 2011
16. Khoury MJ, Coates RJ, Fennell ML, et al: Multilevel research and the challenges of implementing genomic medicine. J Natl Cancer Inst Monogr 2012:112-120, 2012
17. Sackett DL, Straus SE, Richardson WS, et al: Evidence-Based Medicine: How to Practice and Teach EBM. Edinburgh, United Kingdom, Churchill Livingstone, 2000
18. Sackett DL, Rosenberg WM, Gray JAM, et al: Evidence based medicine: What it is and what it isn't. BMJ 312:71, 1996
19. Morton S, Berg A, Levit L, et al: Finding What Works in Health Care: Standards for Systematic Reviews. Washington, DC, National Academies Press, 2011
20. Burge S, Attwood TK, Bateman A, et al: Biocurators and biocuration: Surveying the 21st century challenges. Database (Oxford) 2012:bar059, 2012

21. Moher D, Liberati A, Tetzlaff J, et al: Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. Open Med 3:e123-e130, 2009

22. Higgins JP, Altman DG, Gøtzsche PC, et al: The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 343:d5928, 2011

23. Guyatt GH, Oxman AD, Schünemann HJ, et al: GRADE guidelines: A new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol 64:380-382, 2011

24. Dotson WD, Douglas MP, Kolor K, et al: Prioritizing genomic applications for action by level of evidence: A horizon-scanning method. Clin Pharmacol Ther 95:394-402, 2014

25. Panagiotou OA, Ioannidis JP: Primary study authors of significant studies are more likely to believe that a strong association exists in a heterogeneous meta-analysis compared with methodologists. J Clin Epidemiol 65:740-747, 2012

26. Gøtzsche PC, Ioannidis JP: Content area experts as authors: helpful or harmful for systematic reviews and meta-analyses?. BMJ 345:e7031, 2012

27. Agency for Healthcare Research and Quality: Evidence-based Practice Centers (EPC). Program Overview 2015

28. Greenhalgh J, Dwan K, Boland A, et al: First-line treatment of advanced epidermal growth factor receptor (EGFR) mutation positive non-squamous non-small cell lung cancer. Cochrane Database Syst Rev 5:CD010383, 2016

29. Schuit E, Panagiotou OA, Munafò MR, et al: Pharmacotherapy for smoking cessation: Effects by subgroup defined by genetically informed biomarkers. Cochrane Database Syst Rev 9:CD011823, 2017

30. Schorderet DF: Using OMIM (On-line Mendelian Inheritance in Man) as an expert system in medical genetics. Am J Med Genet 39:278-284, 1991

31. Landrum MJ, Lee JM, Riley GR, et al: ClinVar: Public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 42:D980-D985, 2014

32. Rehm HL, Berg JS, Brooks LD, et al: ClinGen: The clinical genome resource. N Engl J Med 372:2235-2242, 2015

33. My Cancer Genome, 2017

34. Damodaran S, Miya J, Kautto E, et al: Cancer Driver Log (CanDL): Catalog of potentially actionable cancer mutations. J Mol Diagn 17:554-559, 2015

35. Huang L, Fernandes H, Zia H, et al: The cancer precision medicine knowledge base for structured clinical-grade mutations and interpretations. J Am Med Inform Assoc 24:513-519, 2017

36. Griffith M, Spies NC, Krysiak K, et al: CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet 49:170-174, 2017

37. Chakravarty D, Gao J, Phillips SM, et al: OncoKB: A precision oncology knowledge base. JCO Precis Oncol 2017:1-16, 2017

38. Ritter DI, Roychowdhury S, Roy A, et al: Somatic cancer variant curation and harmonization through consensus minimum variant level data. Genome Med 8:117, 2016

39. World Health Organization: ICD-10: International statistical classification of diseases and related health problems: Tenth revision. 2004

40. Stearns MQ, Price C, Spackman KA, et al: SNOMED clinical terms: Overview of the development process and project status. Proc AMIA Symp 662-666, 2001

41. Howe D, Costanzo M, Fey P, et al: Big data: The future of biocuration. Nature 455:47-50, 2008

42. Richards S, Aziz N, Bale S, et al: Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 17:405-424, 2015

43. Hunter JE, Irving SA, Biesecker LG, et al: A standardized, evidence-based protocol to assess clinical actionability of genetic disorders associated with genomic variation. Genet Med 18:1258-1268, 2016

44. Strande NT, Riggs ER, Buchanan AH, et al: Evaluating the clinical validity of gene-disease associations: An evidence-based framework developed by the clinical genome resource. Am J Hum Genet 100:895-906, 2017

45. Elsik CG, Worley KC, Zhang L, et al: Community annotation: Procedures, protocols, and supporting tools. Genome Res 16:1329-1333, 2006

46. UniProt Consortium: UniProt: A hub for protein information. Nucleic Acids Res 43(D1):D204-D212, 2015

47. Edgar R, Domrachev M, Lash AE: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30:207-210, 2002

Appendix

Table A1. Comparison of Key Features of Biocuration Databases

Database	Description	Approach	Curators	Terminal Degrees for Curators
OMIM; www.omim.org[30]	Oldest database on human genes and genetic traits and disorders; focus is on the relationship between genetic variation and phenotypes	Generally a top-down expert review of the published literature on a phenotype and related genes and variations	Researchers with expertise in medical genetics	MD, PhD
UniProt; www.uniprot.org*	Oldest database on protein sequence and functional annotation across organisms, with special emphasis on humans and model organisms; both expert and automatic curation	The SwissProt section contains bottom-up expert review of the published literature on protein function; primary goal is to describe the normal function of proteins, but increasing emphasis placed on annotation of variation and disease association	Researchers with backgrounds in biology, biochemistry, and genetics contributed to the database	PhD, MS
ClinVar; www.ncbi.nlm.nih.gov/clinvar[31†]	Archive of clinically relevant variants and their interpretations	Bottom-up standardized collection of variant interpretations; both case-level and experimental data are included; works with community to set up expert panels; provides clinical significance submitted by the reviewers working with ClinGen to ensure expert curation	Researchers with expertise in medical genetics contributed to the database	PhD
ClinGen; www.clinicalgenome.org[32]	NIH-funded resource dedicated to building an authoritative central resource that defines the clinical relevance of genes and variants	Currently consists of separate disease- and task-specific working groups defining standards, processes, and tooling to curate genes and variants and provide a clinical validity framework; both top down and bottom up; works closely with ClinVar; recent framework proposed for curating cancer variants for clinical utility[38]	Researchers with backgrounds in biology and medical genetics, mainly from institutions involved with the working groups	MD, PhD, MS

(Continued on following page)

Table A1. Comparison of Key Features of Biocuration Databases (Continued)

Database	Description	Approach	Curators	Terminal Degrees for Curators
COSMIC; cancer.sanger.ac.uk‡	Database of somatic mutations in cancer and their impact, including some information on drug sensitivity and resistance; contains expert-curated genes and large-scale genome screening data; also maintains Cancer Gene Census list	Top-down approach to curation; data are compiled and manually curated from scientific literature and other public databases	Researchers with backgrounds in biology, cancer, or genetics contribute to the database	PhD, MS
CIViC; civic.genome.wustl.edu[36]	PO resource describing causal, therapeutic, prognostic, and diagnostic genetic variants	Top-down and bottom-up approaches to literature curation of variants in cancer; assigns evidence levels for clinical utility; community process with curators and editors who review and approve final entries	Researchers with backgrounds in biology and medical genetics	Range of backgrounds, from BS or MS students to MD and PhD
OncoKB; oncokb.org[37]	PO resource containing information about the effects and treatment implications of specific genetic alterations	Top-down approach to curation, where experts curate data from various sources, such as guidelines from the FDA, NCCN, or ASCO, ClinicalTrials.gov, and the scientific literature; provides levels of clinical actionability, with the top level representing FDA approval for the given indication	Researchers and clinical and research fellows at Memorial Sloan Kettering Cancer Center	MD, PhD, MS
MyCancerGenome; www.mycancergenome.org§	PO resource that contains information on tumorigenic mutations and related therapeutic implications, including available clinical trials	Top-down approach to curation, where experts curate data from various sources, such as guidelines from the FDA, NCCN, or ASCO, ClinicalTrials.gov, and scientific literature; text summary includes discussion on whether a particular alteration or variant is prognostic or predictive in given settings	Clinicians and clinical researchers	MD, PhD

NOTE. UniProt is not disease focused; OMIM, ClinVar, and ClinGen are disease agnostic; the remaining databases are specific to cancer.

Abbreviations: ASCO, American Society of Clinical Oncology; CIViC, Clinical Interpretation of Variants in Cancer; ClinGen, Clinical Genome Resource; COSMIC, Catalogue of Somatic Mutations in Cancer; FDA, US Food and Drug Administration; NCCN, National Comprehensive Cancer Network; NIH, National Institutes of Health; OMIM, Online Mendelian Inheritance in Man; PO, precision oncology.

*UniProt Consortium: UniProt: A hub for protein information. Nucleic Acids Res 43:D204-D212, 2015.

†Landrum MJ, Lee JM, Benson M, et al: ClinVar: Public archive of interpretations of clinically relevant variants. Nucleic Acids Res 44:D862-D868, 2016

‡Forbes SA, Beare D, Boutselakis H, et al: COSMIC: Somatic cancer genetics at high-resolution. Nucleic Acids Res 45:D777-D783, 2017.

§Taylor AD, Micheel CM, Anderson IA, et al: The path(way) less traveled: A pathway-oriented approach to providing information about precision cancer medicine on My Cancer Genome. Transl Oncol 9:163-165, 2016.

Received 12 August 2013
Revised 5 January 2014
Accepted 7 January 2014

Domain knowledge and data quality perceptions in genome curation work

Hong Huang
School of Information, University of South Florida, Florida, USA

Abstract

Purpose – The purpose of this paper is to understand genomics scientists' perceptions in data quality assurances based on their domain knowledge.

Design/methodology/approach – The study used a survey method to collect responses from 149 genomics scientists grouped by domain knowledge. They ranked the top-five quality criteria based on hypothetical curation scenarios. The results were compared using χ^2 test.

Findings – Scientists with domain knowledge of biology, bioinformatics, and computational science did not reach a consensus in ranking data quality criteria. Findings showed that biologists cared more about curated data that can be concise and traceable. They were also concerned about skills dealing with information overloading. Computational scientists on the other hand value making curation understandable. They paid more attention to the specific skills for data wrangling.

Originality/value – This study takes a new approach in comparing the data quality perceptions for scientists across different domains of knowledge. Few studies have been able to synthesize models to interpret data quality perception across domains. The findings may help develop data quality assurance policies, training seminars, and maximize the efficiency of genome data management.

Keywords Behaviour, Assessment, Curation, Domain knowledge, Genome

Paper type Research paper

Introduction

The proliferation of heterogeneous genomic data types represents the diverse concepts of biology (Sanderson, 2011; Wu *et al.*, 2010; Yang *et al.*, 2011). Genome curation is the process of digitizing and integrating disparate pieces of genomic data and their related literatures to facilitate the sharing of genomic knowledge (Reed *et al.*, 2006). The genome curation process can be facilitated by using standardized terminologies and metadata schemas (MacMullen and Denn, 2005; Pagani *et al.*, 2012; Willis *et al.*, 2012). There are well established terminologies and metadata standards in biosciences for describing data-types, protocols used in experiments, and gene ontology for molecular functions (Leonelli *et al.*, 2011; Mayor and Robinson, 2014). It is a complex process that requires multidisciplinary knowledge, pertinent work experience, and skills relevant to the effective execution of multi-faceted curation operations (Burkhardt *et al.*, 2006). Thus, genomic research has become a data rich domain requiring not only effective methods to process, interpret, and reuse genomic data (Salimi and Vita, 2006; Samuel *et al.*, 2008), but extensive knowledge of the fields of biology, bioinformatics, and computational science.

Scientists working on genome curation require domain knowledge in areas such as biology, bioinformatics and computational science. Scientists conducting genome curation generally possess either PhDs or Masters degrees in biology, bioinformatics, computer science, or other related disciplines (Burge *et al.*, 2012). Wet-lab research experience in biochemistry and molecular biology contributes meaningfully to their collective ability to determine and select the desired information resources that can help

Journal of Documentation
Vol. 71 No. 1, 2015
pp. 116-142

0022-0418
DOI 10.1108/JD-08-2013-0104

their curation or annotation work (Burge *et al.*, 2012). In certain cases, subject expertise or domain knowledge is essential to ensure acceptable upstream phases of genomic data management and planning (Bentley, 2006). Genomics scientists with a wide array of experience, participate in comprehensive training and workshops in order to improve their curation skills (Sanderson, 2011; Shimoyama *et al.*, 2009). They also consult the curation manual regularly to ensure that they follow curation standards in identifying data elements, assigning nomenclature, and annotating genomic-related data with biological information (Samuel *et al.*, 2008).

It has been found that genomics scientists have shared certain requirements for data quality, leading to the development of a general data quality model for genome curation (Huang *et al.*, 2012).Within genome curation, the context for both information use and information operation is complicated. As a result of varieties of domain knowledge exists among genomic scientists, the respective complexities of domain knowledge and work experience might consequently affect scientists' decision making. The relationship between domain knowledge types for genome curation and data quality assurance activities remains unknown. Scientists from different domains and backgrounds could make conflicting data quality decisions when assigning the same genome curation tasks. It could result in the current gap in understanding of the curation problems associated with data quality assurance when different domain knowledge is exchanged among biology, bioinformatics, and computational science.

The purpose of this study is to understand the relationship between different types of domain knowledge and scientists' data quality requirements. Specifically, the respective performances of three different user groups, who possess domain knowledge in the fields of computational science, bioinformatics, and biology, will be examined in order to identify their perceptions of data quality requirements. The findings could benefit the development of domain sensitive data quality and skill models for genomic research communities, yielding both improved resource integration and more cost-effective collaborative solutions.

Literature review

Scientists conducting genome curation work have been trained in their disciplinary knowledge (e.g. biology, computer science) at the post-graduate degree level or higher. Biological research has progressed to an intensive data process and evaluation using multiple data mining tools. The data-driven approach has become a common research practice for scientists (Reed *et al.*, 2006; Goth, 2012). Data curation and manipulation tools need to be customized by scientists to fit into a specific biological context (Lathe *et al.*, 2008; Huang *et al.*, 2011; Pruitt *et al.*, 2012). Biologists also need data analysis support from computational scientists to process the massive data sets produced through their research. The task is not easy because the traditions and cultures of these domains are not the same (Wooley and Lin, 2005). Genomics scientists need much closer scrutiny to explicate the characteristics of domain knowledge in both biology and computer science. It is through such scrutiny that they can adopt effective practices for data quality assurance and data exchange among distinct disciplines.

Domain knowledge in genome curation

Genomics research has grown and changed rapidly. Genomic data curation originally started as sequence analysis only (Reed *et al.*, 2006). It has since incorporated a wide variety of data processes and analysis such as genome-wide association studies, micro

arrays, protein-protein interactions, and literature text-mining (Cole and Bawden, 1996; Bartlett and Toms, 2005; Ioannidis and Khoury, 2011; Lathe *et al.*, 2008; Sanderson, 2011; Shachak and Fine, 2008). Curating genomic data is a highly interdisciplinary process requiring scientists to have diverse skills.

Domain knowledge can be defined as the degree of familiarity with a particular domain or subject area (Allen, 1991; Ju, 2007; Wildemuth, 2004). It encompasses declarative knowledge (knowing what), procedural knowledge (knowing how), and conditional knowledge (knowing when and where), (Alexander, 1992; Hjørland and Albrechtsen, 1995). Domain and discipline knowledge seem to "fall along a continuum that is defined by both external and internal factors" (Alexander, 1992, p. 36). Relationships within a domain, the rules of that domain, and its historical context all need to be considered to embrace the complete and meaningful domain knowledge of a discipline (Hjørland and Albrechtsen, 1995). Genome science is an interdisciplinary field that requires collaborative work with both biologists and computational scientists. Wooley and Lin (2005) distinguish biologists from computational scientists in their research goals and working practices (see Table I).

Computational science develops algorithms and software tools to support data retrieval, organization and analysis (Fenstermacher, 2005). However, there are distinct sets of rules for data configuration and operations between biology and computer science (Wooley and Lin, 2005). Biologists are particularly interested in seeking "signal in the noise of their experimental data" (Wooley and Lin, 2005, p. 367). Since biological research is driven by experiment and observation, its goals consist of finding solutions to individual and specific problems. In contrast, computational scientists are trained to "search for boundary conditions and constraints" (Wooley and Lin, 2005, p. 367). Computational science research is driven by analytical methods and techniques, and its research goals are the development of solutions that can solve many problems. Computational scientists who work with biological data are trained to "take categorical statements literally, whereas biologists use them informally" (Wooley and Lin, 2005, p. 367).

Table I.
Examples of working objectives and practices in biology and computational science

	Biology	Computational science
Working objectives	Understand the mechanism of development for living organisms, and then use that understanding to determine examples of application areas for biological data	Identify the unknown patterns within massive biological data sets
	Seek signals in the noise of their experimental data	Search for boundary conditions and constraints
	Provide solutions to individual and specific problems	Develop universal solutions to solve many problems
Working practices	Research is driven by experiment and observation	Research is driven by analytical methods and techniques
	Question the mathematical soundness of their approach by providing exceptions to their cases	May underestimate the complexity of the biological problems, oversimplify biological models and give out universal statements that fall short of expected exceptions
	Limited freedom to establish rules	Open to the establishment of their own rules for developing algorithms
	Use categorical statements informally	Take categorical statements literally

Sources: Wooley and Lin (2005)

Because of the constraints imposed by nature, biology has limited freedom to establish rules. These constraints are consistent with the rules applied to the biological phenomena. In contrast, computer science is open to the establishment of one's own rules provided that doing so allows sense to be made of the algorithm (Wooley and Lin, 2005). Biologists might focus on understanding the mechanism of development for living organisms, and then use that understanding to determine examples of application areas for biological data (Wooley and Lin, 2005). In contrast, computational scientists are data scientists. They are more engaged in attempts to identify the unknown patterns within massive data sets (Wooley and Lin, 2005). As programmers, computational scientists could easily underestimate the complexity of the biological problems, and therefore both oversimplify biological models as well as give out universal statements that fall short of expected exceptions (Wooley and Lin, 2005). However, biologists, particularly those untrained in quantitative sciences, always question the mathematical soundness of their approach by providing exceptions to their cases (Wooley and Lin, 2005). During the genome curation process, both the biologist and the computational scientists collaborate with each other. During this process, however, they may experience conflicts and disagreements in defining curation roles and thus yield contested interpretation of curated data.

Previous research indicates that scientists' domain knowledge affects their information seeking behavior and their interactions with information systems and software tools (Brown, 2003; Hemminger *et al.*, 2007; Vibert *et al.*, 2009; Wu *et al.*, 2012). It can be assumed that biologists have a high level of declarative knowledge of biology, whereas computational scientists have a high level of procedural knowledge of computer systems. Although both biologists and computational scientists might be expected to know how to use computer programs or curation tools, computational scientists probably enjoy a broader knowledge of tools and programs. However, when a biologist interacts with a new curation tool, s/he holds the advantage over a computational scientist of knowing the particular semantics (the words or terminologies about biological concepts) used in that program (Bartlett and Toms, 2005; Chilana *et al.*, 2009). In other tasks such as accurately predicting the options available in a generic help menu or in the functions of menu interface designed for automatic genome annotation systems, the biologist may be at a disadvantage compared to the computational scientist (Chilana *et al.*, 2009; Shachak and Fine, 2008).

Domain knowledge affects scientists' decisions in the determination of data processing strategies, data-quality assurance activities, analytic tools selection, and result evaluation (Chilana *et al.*, 2009; Ju, 2007; Vibert *et al.*, 2007, 2009; Wu *et al.*, 2012). The scientists with biology domain knowledge could easily find the exceptions or special cases (Wooley and Lin, 2005) for which annotation tools and guidelines might not yet be available. Similarly, computational scientists can benefit from the wet-lab experiences of biologists to develop both complex software tools and standardized workflows (Chilana *et al.*, 2009). Scientists need to remain open to explore new research opportunities in a typical domain as an "outsider," and develop strategies for exploring and translating information from unfamiliar domains to manage their interdisciplinary information work (Palmer and Neumann, 2002). Development of a comprehensive data curation model can help yield the high-quality curation products that both biology and computational science require. Such a data model encourages experts from two domains to work closely with each other thereby reducing domain crossing barriers while merging knowledge across disciplinary boundaries (Haythornthwaite, 2006; Klein *et al.*, 1997).

Data quality and domain knowledge

There are different working domains and scholarly contexts through which data quality can be both operationalized and defined. It has been argued that data quality as a concept is contextual and must be evaluated within the context (Strong *et al.*, 1997; Stvilia *et al.*, 2007). An aspect of a DQ concept is defined as a DQ dimension (Huang *et al.*, 2012; Stvilia *et al.*, 2007; Wang and Strong, 1996). Several studies have assessed specific DQ dimensions in different domains. One study explored progress in the accuracy assessments of automated genome curation tasks (Brent, 2008), whereas another examined in an online interactive community, for patterns of credibility (Lankes, 2008). Wang and Strong (1996, p. 6) provided a definition for quality, describing it as "fitness for use." This indicated the importance of defining data quality within context of use (Strong *et al.*, 1997; Stvilia *et al.*, 2007). The need to comprehend the extent to which user satisfaction is realized has the potential to characterize data quality within a particular context (Evans and Lindsay, 2005; Huang *et al.*, 2012).

According to research, sets of DQ dimensions that have been determined to be important, include those pertaining to gene-ontology curation behaviors (MacMullen, 2006), online scholarly information (Rieh, 2002), and consumer health information (Frické and Fallis, 2004; Stvilia *et al.*, 2007). Genomics scientists suggest that trust-related dimensions such as Unbiased and Believability are important in genome curation when they indirectly assess the quality of curation data (Huang *et al.*, 2012). Data quality aspects related to trust help scientists gauge the degree of confidence they can have. Studies have shown that domain knowledge could facilitate researchers in evaluating the trustworthiness of reference sources (Vibert *et al.*, 2009). Data standards, metadata schemas, and curated databases were developed to facilitate the accessibility of disparate genomic data sets (Barrett *et al.*, 2012; Willis *et al.*, 2012). DQ models were developed to describe and capture the overall value structure and the context for DQ for a genome curation community (Huang *et al.*, 2012), a Wikipedia community (Stvilia *et al.*, 2007) and online health information consumers (Stvilia *et al.*, 2008).

Lee and Strong (2003) have argued that three knowledge modes are related to data quality dimensions. According to Lee and Strong (2003), the declarative, or knowing-what, may be defined as understanding the activities through which the data production processes are realized. Procedure, or knowing how, is defined as understanding procedures needed to respond to known DQ difficulties and obstacles (Lee and Strong, 2003). Knowing-why is defined as contextual knowledge that can formulate the questions to understand related purposes and the ability to analyze underlying principles (Lee and Strong, 2003). During the data process, it has been found that the prioritization of DQ dimensions differs among users with varying knowledge modes (Lee and Strong, 2003). The genome curation community in fact requires a set of DQ skills to guarantee data quality itself. Genome curation work requires excellent written and verbal communication skills to facilitate the acquisition and description of genomics data. Knowledge in biology and/or bioinformatics also helps to evaluate quality control of experimental data. Genome curation work is data-driven; much of the scientists' time is spent on data wrangling or "munging," ie. dealing with the large scale of genomic data for data preprocessing, integration, data cleaning and validation (Heer and Kandel, 2012; Reed *et al.*, 2006). Through a survey of DQ professionals who hold a series of professional employment positions, Chung *et al.* in 2002 created a practical educational framework. The framework described three useful DQ categories, each one pertaining to a particular set of capabilities of DQ skills, specifically technical, adaptive, and interpretive.

Domain experts obtain domain-specific knowledge, work-related experience, and trainings. This experience or knowledge can also support data-quality related activities and allow domain experts to make greater use of data-quality information than those without related knowledge (Fisher *et al.*, 2003). Users with experience or domain knowledge might be sensitive in detecting both errors and missing data (Klein *et al.*, 1997; Sanbonmatsu *et al.*, 1992), adaptive in using contextual and relevant information (Sanbonmatsu *et al.*, 1992; Payne *et al.*, 1993), and proficient in organizing information (Mackay and Elam, 1992). Domain knowledge could guide users to effectively test the validity of their discovered knowledge (Owrang and Grupe, 1996). Domain knowledge could also improve the performance of information seeking (Marchionini *et al.*, 1993; Tabatabai and Shore, 2005; Vibert *et al.*, 2007). Users with excellent domain knowledge may have greater accessibility to desired information, more flexibility to handle relevant information, and better contribution to knowledge representation (Rouet *et al.*, 1997; Vibert *et al.*, 2009). The processing of extensive knowledge of information sources in their disciplines aids domain experts in the evaluation of both the usefulness and trustworthiness of documents (Vibert *et al.*, 2009).

Differences in knowledge and experience across domains also create barriers to a consensus in work activities or processes in an interdisciplinary collaborative work environment (Wooley and Lin, 2005). Paradigms in a particular domain can be referred to as concrete problem solutions, procedures of experiments, and theoretical models shared by the scientists in a community (Kuhn, 1974; Eysenck, 1991). However, counting on paradigms to formalize scientific thinking might possibly limit the development and evolution of a discipline (Watt, 2000). In addition, prior experience or knowledge is not always a positive (Fisher *et al.*, 2003). For example, experience or work knowledge might affect users' perceptions and expectations for data quality (Klein *et al.*, 1997), and may cut off the decision process unacceptably early (Dukerich and Nichols, 1991). Sometimes, users with sufficient knowledge might show less attention to related information (e.g. data quality information) than those who do not have such knowledge (Yates *et al.*, 1991). They might also be more inclined to perform tasks less accurately than users without prior experience (Gilliland *et al.*, 1994). Genome data curation is performed by scientists with different domain knowledge and skills. Domain knowledge differences in genomics scientists could influence the beliefs and expectations of data quality assurance activities for genome-curation specific annotation tasks and activities.

Research questions

This was an exploratory study. It sought to understand the relationship between perception of DQ dimensions and skills and domain knowledge among genomics scientists. Specifically, the study investigated the following two research questions:

RQ1. How do genomics scientists with different domain knowledge of genomic curation processes prioritize DQ dimensions?

This question is explored through comparing survey rankings of DQ dimensions among biologists, bioinformaticians, and computational scientists in genome curation.

RQ2. How do genomics scientists with different domain knowledge of genomic curation processes prioritize DQ skills?

This question is investigated by comparing the survey rankings of DQ skills among biologists, bioinformaticians, and computational scientists in genome curation.

Methods

The study collected and analyzed survey data. The survey questions were collected and modified from the previous DQ dimensions and skills items found in the literature (Chung *et al.*, 2002; Wang and Strong, 1996). Survey participants were genomics scientists who had published journal articles related to genome annotation, curation, and genomic research. Participants were given two scenarios that represented and conceptualized genome curation activities. These scenarios were developed by using scenario-based task analysis (Carroll, 1997; Diaper, 2004; Go and Carroll, 2004; Huang *et al.*, 2012). Participants were provided the same set of written requirements for genome curation that can be used for understanding user perception (see Appendix 1). Scientists thus can perceive the data quality requirements provided by a common set of curation tasks as scenarios. The first scenario asked scientists to pick the top five DQ dimensions, from a total 17 DQ dimensions; the second scenario asked for the ranking of the top five DQ skills, from a total of 17 DQ skills (Table II and Table III). In addition, the subjects were asked to open-ended comments on the clarity and comprehensibility of the survey questions, as well as additional concerns about data quality or skills in genome curation. The 149 survey respondents were further grouped by their domain knowledge, specifically biology, computational science, and bioinformatics.

Groups	Data-quality dimensions
Accuracy	Accuracy: sequence records are correct and free of error
	Unbiased: sequence records are unbiased and objective
	Believability: sequence records are regarded as credible and believable
Accessibility	Accessibility: sequence records are easily and quickly retrievable for access
	Traceability: the derivation history of the sequence records is documented and traceable
	Appropriate amount of information: the volume of the sequence records is appropriate for this scenario
Usefulness	Interpretability: sequence records are in appropriate languages, symbols, and units, and the definitions are clear for interpretation
	Understandability: sequence records are easily understandable
	Ease of manipulation: sequence records are easy to manipulate and make it easy to carry out various tasks described in this scenario
	Consistency: sequence records are presented in a consistent format
	Value-added: sequence records contain additional annotations from the tasks in this scenario and these annotations are beneficial and add value
Relevancy	Relevancy: sequence records contain information relevant to the scenario
	Concise representation: sequence records are concisely represented
	Completeness: annotated sequence records are not missing and are fully annotated according to the steps described in this scenario
	Up-to-date: sequence records are sufficiently up-to-date for this scenario
	Reputation: sequence records are highly regarded and reputable in terms of their source or content

Table II.
List of data quality dimensions for top-five rankings and their categories

Note: Lists of data-quality dimensions and their groupings based on previously reported data quality dimensions and skills models
Sources: Chung *et al.* (2002), Wang and Strong (1996), Huang *et al.* (2012)

Groups	Data-quality skills
Adaptive skills	User requirement: ability to translate subjective user requirements for data quality into objective technical specification (such as use of quality function deployment)
	Data entry improvement: skills and ability to analyze and improve the data entry process in order to maintain data quality
	Organization policies: ability to establish and maintain organizational policies and rules for data quality management
	Change process: ability to manage the change process/transitions resulting from the data quality management project
	Data quality cost/benefit: skills and ability to conduct cost/benefit analysis of data quality management
	Information overload: understanding the information overload that managers often face and ability to reduce information overload
Interpretative skills	Data error detection: ability to detect and correct errors in databases
	Software tools: experience and ability to use diverse commercially available data quality software packages
DQ literacy skills	Data quality dimensions: quality dimensions are concepts/"virtues" that define data quality. Data quality dimension skills are the ability to define and describe diverse dimensions of data quality (such as relevancy, believability, accessibility, ease of understanding)
	Data quality measurement: data quality measurement is an operationalization of a data quality dimension. Data quality measurement skills are the ability of assessing the variation along the dimension
	Data quality implication: understanding pervasiveness of data quality problems and their potential impacts
Technical skills	Data quality audit: ability to conduct data quality auditing (formal review, examination, and verification of data quality)
	Statistical techniques: ability to apply statistical techniques to manage and control data quality
	Data mining skills: data mining and knowledge discovery skills for analyzing data in a data warehouse
	Data warehouse setup: ability to integrate multiple databases into an integrated data warehouse
	Analytic models: ability to apply diverse analytic models (such as regression model and multidimensional model) for data analysis
	Structural query language (SQL): skills and ability to apply SQL to estimate the accuracy of data

Note: Lists of data-quality skills and their groupings based on previously reported data quality dimensions and skills models
Sources: Chung *et al.*, (2002), Wang and Strong (1996), Huang *et al.* (2012)

Table III.
List of data-quality skills for top-five rankings and their categories

Scientists who selected trainings in both biology and computer science related disciplines were grouped as "Bioinformatics" ($n = 38$). Bioinformaticians have knowledge proficiency in both biology and computer science domains. Additionally, those who chose computer science and related disciplines were grouped as "Computational Science" ($n = 24$). Last, scientists with biology training and wet-lab experience were grouped as "Biology" ($n = 87$). For curation experience, the majority (90 percent of the participants) of the scientists in this study had one year or more work experience in genome curation, and 40 percent had more than five years' experience. With regard to age, 88 percent of the participants were between 30 and 40 years old.

Distribution and collection of the survey was conducted online through the Qualtrics software (www.qualtrics.com). The survey data was analyzed with STATA 11 software (College Station, Texas, USA) to perform descriptive statistics and χ^2 analysis. The author computed the percentiles for the occurrences for each of the 17 DQ dimensions and DQ skills being ranked by the users as the top five DQ dimensions or skills. Next, the computed percentiles of each DQ dimension and skill were ranked from the largest to the smallest, and then the cumulative percentage for each DQ dimension and skill were also calculated (see Appendix 2 and 3). The cumulative percentage for each DQ dimension or skill was calculated as follows:

$$Yi = \sum_{i=1}^{k} Xi \Big/ \sum_{j=1}^{N} Xj$$

X represents the percentile value for a DQ dimension or DQ skill for the number of top-five ranking occurrences divided by the total top five ranking occurrences. Yi is defined as the cumulative percentage for the ith ranking of DQ dimensions or skills accumulated from the percentiles from the first DQ dimension or skill ranking to the ith. i takes values from 1 to k. k is the number of DQ dimensions or skills accumulated from the fisrt to ith rank. The value of jth is from 1 to N. N is the total number of DQ dimensions ($N = 17$) or skills ($N = 17$). For that reason, the value of Y for the last accumulated ranking (*17*th) for DQ dimensions or skills is 100 percent (see Appendix 2 and 3). Only those DQ dimensions or skills with cumulative rankings less than 90 percent were kept as those greater than 90 percent only count for a trivial portion – specifically, less than 10 percent of total occurrences in top five ranking chosen by the users – and can be ignored.

The selected DQ dimensions and skills within the top 90 percent accumulated ranking lists were further grouped into categories based on previous reported data quality dimensions and skills models (Chung *et al.*, 2002; Wang and Strong, 1996; Huang *et al.*, 2012) as represented in Table II and III. Finally, the aggregated percentage was computed for each category, for both DQ dimensions and skills models, by adding up the percentile of each dimension or skill in a category. The researchers then ranked these categories in decreasing order based on their aggregated percentages (Appendix 2 and 3).

Findings

χ^2 analysis of the genome curation survey results for the top-five DQ dimensions and skills selections and rankings found differences in priorities of specific DQ skills and dimensions. Some of these differences were statistically significant. Those DQ dimensions and skills that were affected by domain knowledge were identified. Furthermore, there are specific DQ trade-offs for a typical group of DQ dimensions and skills found in different user groups, particularly among computational scientists and biologists. DQ trade-offs occurred when the DQ expectations of scientists did not match the actual needs in the domain. Data curation models or policies can in fact be defined more specifically to meet the domain dependent needs, suggesting that new curation procedures and data standards need to be developed in order to accommodate different requirements among users.

The descriptive statistical analysis of the survey data for the occurrences of each DQ dimension revealed the top-five most important DQ dimensions for each group of

scientists, ranked from highest to lowest. Table IV and V showed the descriptive summary of the rankings for all the DQ dimensions and skills in different domain experts. Particularly, the statistical significant ones and their χ^2 values were bold/italic, and cells of the top five rankings were also highlighted for each group. The five most important dimensions for computational scientists were: accuracy, accessibility, completeness, understandability, and appropriate amount of information. According to biologists, the five most important DQ dimensions were: accuracy, accessibility, completeness, believability, and up-to-date. Last, bioinformaticians ranked the top-five DQ dimensions as: accuracy, accessibility, completeness, believability, and interpretability. It is worth noting that for all three groups accuracy, accessibility, and Completeness were among the most important DQ dimensions. Interestingly, computational scientists did not rank Believability as one of the top five, but both biologists and bioinformaticians did. In contrast, computational scientists ranked understandability as of particular importance. Biologists were interested in believability and currency ("Up-to-date") and bioinformaticians cared more about interpretability.

χ^2 analysis (Table IV) found several significant differences in data quality perceptions among scientists with different domain knowledge. Compared to biologists and bioinformaticians, computational scientists held a higher expectation in understandability and a stronger need for concise representation. Bioinformaticians expressed a particular interest in interpretability. Unlike computational scientists, both biologists and bioinformaticians ranked believability as one of the five most important dimensions. Biologists also ranked traceability higher than other two groups (Table IV).

As for DQ skills (Table V), all three user groups shared the belief that data error detection, data mining skills, DQ quality measurement, and statistical techniques were very important DQ skills for genome curation work. Biologists have a stronger need for two DQ literacy skills: DQ measurement, DQ implication. Bioinformaticians care about DQ literacy skills specifically DQ measurement, and DQ dimensions (Table V). While there are some shared preferences between groups, the results also indicated as well, that the ranking of skills varied. Computational scientists ranked from highest to lowest, what they felt to be the most important DQ skills as data-error detection, DQ measurement, statistical techniques, data mining skills, and DQ implication. Among biologists, the top five DQ skills were ranked from highest to lowest as data-error detection, DQ measurement, data mining skills, statistical techniques, and DQ implication. And bioinformaticians ranked the top five most important to least as DQ error detection, data mining skills, statistical techniques, DQ dimensions, and DQ measurement.

Among these groups, data quality error detection was found to be the most important skill when performing annotation work within the genome annotation context. When looking at the ranking patterns among biologists, computational scientists, and bioinformatics, the importance rankings include data quality literacy skills as well as interpretative skills. Interestingly, importance rankings as demonstrated in Table V, indicate a strong demand by computational scientists for statistics techniques.

χ^2 analysis results also suggest that there is a stronger preference for data warehouse setup and information overloading skills for biologists than computational scientists and computational biologists care more about structure query language (SQL) than the other two groups. It is worth noting however, that bioinformaticians, as indicated in Table V, have higher expectations regarding Data mining skills than do biologists.

In regard to domain knowledge-based differences evidenced in the rankings of four DQ dimension categories (Figure 1), all three user groups regarded the accuracy group

Table IV.
Rankings of DQ dimensions based on the domain knowledge

Attribute	Computation ($n = 24$) Mean rank	Computation ($n = 24$) Ranked by	Biology ($n = 87$) Mean rank	Biology ($n = 87$) Ranked by	Bioinformatics ($n = 38$) Mean rank	Bioinformatics ($n = 38$) Ranked by	χ^2 (Comp. vs Bioinf.)	χ^2 (Comp. vs Biol.)	χ^2 (Biol. vs Bioinf.)
Accessibility	1.7	17 (70.8%)	2.6	58 (66.7%)	1.9	22 (57.9%)	1.055	0.149	0.883
Accuracy	2.5	17 (70.8%)	1.9	64 (73.6%)	1.7	28 (73.7%)	0.060	0.071	0.0002
Appropriate amount of information	2.0	7 (29.2%)	2.3	26 (29.9%)	3.3	10 (26.3%)	0.060	0.005	0.164
Believability	2.8	7 (29.2%)	2.9	34 (39.1%)	2.5	17 (44.7%)	1.503	0.794	0.350
Completeness	4.0	10 (41.7%)	3.0	43 (49.4%)	2.9	20 (52.6%)	0.708	0.454	0.109
Concise representation	3.7	5 (20.8%)	4.7	4 (4.6%)	4.3	3 (7.9%)	2.191	*6.655*	0.544
Consistent representation	2.5	7 (29.2%)	2.8	30 (34.5%)	3.7	14 (36.8%)	0.387	0.239	0.065
Ease of manipulation	2.7	5 (20.8%)	3.3	23 (26.4%)	4.0	9 (23.7%)	0.068	0.313	0.105
Interpretability	4.3	7 (29.2%)	3.5	14 (16.1%)	4.1	16 (42.1%)	1.055	2.096	*9.813*
Relevance	4.0	3 (12.5%)	3.8	7 (8.0%)	3.0	4 (10.5%)	0.057	0.455	0.203
Reputation	2.0	2 (8.3%)	3.3	6 (6.9%)	3.5	4 (10.5%)	0.081	0.058	0.473
Security	0.0	0 (0.0%)	3.0	7 (8.0%)	4.4	2 (5.3%)	1.305	2.061	0.307
Traceability	2.0	2 (8.0%)	3.8	24 (27.6%)	4.1	8 (21.1%)	1.759	*3.887*	0.593
Unbiased	3.3	7 (29.2%)	3.1	21 (24.1%)	3.3	10 (26.3%)	0.060	0.252	0.067
Understandability	3.8	10 (41.7%)	4.1	13 (14.9%)	3.8	7 (18.4%)	*3.994*	*8.178*	0.238
Up-to-date	3.8	7 (29.2%)	4.1	33 (37.9%)	4.0	9 (23.7%)	0.231	0.627	2.406
Value added	5.0	2 (8.3%)	4.0	7 (8.0%)	4.0	5 (13.2%)	0.342	0.0021	0.796

Notes: Italics: χ^2 scores were statistically significant ($p < 0.05$). Top five DQ skills for each group have the cell highlighted

Table V.
Rankings of DQ skills based on the domain knowledge

Attribute	Computation ($n = 24$) Mean rank	Computation ($n = 24$) Ranked by	Bioinformatics ($n = 38$) Mean rank	Bioinformatics ($n = 38$) Ranked by	Biology ($n = 87$) Mean rank	Biology ($n = 87$) Ranked by	χ^2 (Comp vs Bioinf)	χ^2 (Comp vs Biol)	χ^2 (Bioinf vs Biol)
Analytic models	2.5	3 (12.5%)	3.7	6 (15.8%)	4.2	19 (21.8%)	0.128	1.032	0.605
Change process	3.0	2 (8.3%)	4.5	2 (5.3%)	4.1	11 (12.6%)	0.230	0.338	1.546
Data mining skills	3.3	10 (41.7%)	3.3	23 (60.5%)	2.5	34 (39.1%)	2.102	0.053	*4.904*
Data-entry improvement	3.5	7 (29.2%)	3.4	7 (18.4%)	3.0	27 (31%)	0.972	0.031	2.125
Data-error detection	1.7	15 (62.5%)	2.8	30 (78.9%)	2.6	58 (66.7%)	1.999	0.145	1.914
Data-quality audit	3.4	9 (37.5%)	3.3	7 (18.4%)	3.6	24 (27.6%)	2.797	0.885	1.191
Data-quality cost/benefit	0.0	0 (0.0%)	3.7	3 (7.9%)	4.3	6 (6.9%)	2.000	1.750	0.039
Data-quality dimensions	2.0	7 (29.2%)	1.7	15 (39.5%)	2.0	29 (33.3%)	0.683	0.149	0.437
Data-quality implication	2.6	9 (37.5%)	2.4	8 (21.1%)	2.7	30 (34.5%)	1.999	0.075	2.255
Data-quality measurement	3.2	10 (41.7%)	2.4	14 (36.8%)	1.9	34 (39.1%)	0.144	0.053	0.056
Data-ware house set-up	0.0	0 (0.0%)	3.7	9 (23.7%)	3.4	16 (18.4%)	*6.649*	*5.157*	0.463
Information overload	1.7	6 (25%)	2.5	2 (5.3%)	4.6	7 (8.0%)	*5.009*	*5.229*	0.307
Organization policies	3.7	5 (20.8%)	2.7	10 (26.3%)	3.2	16 (18.4%)	0.241	0.073	1.008
Software tools	4.6	9 (37.5%)	3.7	12 (31.6%)	3.5	23 (26.4%)	0.230	1.122	0.347
Statistical techniques	3.3	10 (41.7%)	3.3	17 (44.7%)	3.7	30 (34.5%)	0.056	0.421	1.185
Structure query language	1.0	2 (8.3%)	4.0	1 (2.6%)	0.0	0 (0.0%)	1.039	*7.383*	2.308
User requirement	4.0	7 (29.2%)	2.8	5 (13.2%)	3.4	23 (26.4%)	2.415	0.071	2.683

Notes: Italics: χ^2 scores were statistically significant ($p < 0.05$). Top five DQ skills for each group have the cell highlighted

as the primary DQ concerns in genome curation work. Rankings also indicated that biologists care more about the data accessibility issues than the other two groups. Both computational scientists and bioinformaticians care more about usefulness of current curation than the biologists.

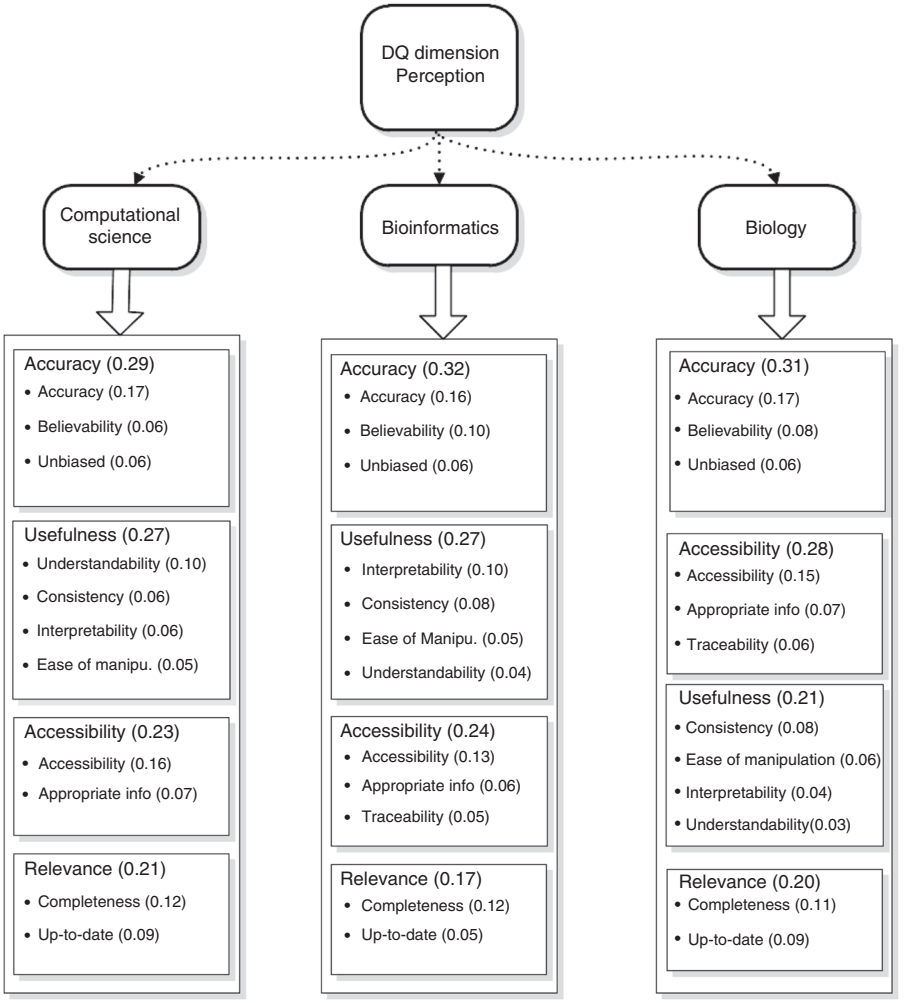
The study also compared the rankings of the four DQ skills categories by participants with different domain knowledge based on previous data skills models (Table III). The results are shown in Figure 2. The findings indicated that computational scientists consider adaptive skills more important for genome curation work than did the other two user groups. All three user groups, however, regarded technical skills as important for dealing with genome curation.

Discussion

This study determined that scientists with different domain knowledge prioritize DQ dimensions and/or skills differently. The first research question focussed on DQ dimension perception gaps among users with different domain knowledge. Significant perception differences were found among all three groups in the categories of relevant information (concise representation) and useful information (interpretability, understandability) (Figure 1). Believability was indicated among the top-five DQ dimensions for both biologists and bioinformaticians, but not for computational scientists. Users with different domain knowledge also assigned different priorities among the DQ skills requirements. These differences are observed among the technical skills; specifically data mining skills, data warehouse set-up, and SQL. Because users with different domain knowledge held specific sets of prioritized DQ dimensions and skills requirements, the contextualized data quality models were defined based on the domain knowledge of the users.

Domain knowledge and DQ dimensions perception

Knowing-what knowledge required scientists to define the genome related biological research questions/curation goals. This requires theoretical biological knowledge to understand what to do with genomic data. Knowing-how, formally known as procedure knowledge, refers to the ability to carry out a task through sequential procedures, such as running the sequential tasks for genome sequences analysis. Procedural knowledge required scientists to develop automatic annotation tools and procedures for the support of genome curation work. Having obtained this knowledge, scientists may then focus on the development of a practical solution to the curation problem related to genome curation work. Knowing why knowledge is defined as the understanding of the reasons and principles underlying the work practice (Lee and Strong, 2003). In genome curation work, scientists who hold knowledge of both biology and computer science are at a greater advantage of understanding the purpose of the curation work in both genome curation activities and procedures. That is why genomics scientists require cross-domain knowledge/skills both in biology and computational science for curation work. Biologists are trained by means of wet-lab experiments, making an extra effort to understand the context of data derived by an unfamiliar technique. Learning new lab techniques takes many years to master, making high-quality data appraisal difficult. Biologists are most likely interested in interpretation of the curation data through their knowledge of biology, in the help of computational scientists or programmers in data interpretation, and in explanation of the curation requirements for software development. Biologists examine authoritative data sources and evaluate their



Note: Only those with cumulative rankings of less than 90 percent were kept (see Appendix 2)

Figure 1.
Domain knowledge based DQ dimension priorities

annotation. They therefore care more about the believability of the data sources. Biologists also pay attention to the exceptions/special cases of biological knowledge (Wooley and Lin, 2005). It is important to develop genome curation systems that enable the trace function to fully capture the update of curated biological knowledge for data access and preservation (Shimoyama *et al.*, 2009). Computational scientists or programmers, however, obtain training in both the procedures and data mining protocols of the genome curation process and related issues in data management. They focus more on developing a technical and practical solution to a biological problem in data curation.

Depending on the curation problems and selected approaches, scientists might experience a mismatch of their understanding in a single aspect of data quality for their

curation needs. For instance, accessibility barriers might be perceived differently among users. Some scientists might think certain genomic data simply physically unavailable rather than inaccessible. However, other scientists might interpret the barriers as being technical, based on the following reasons: the coded data may be barely interpreted; data may be represented in different formats which are unrecognized; and a large volume of data is in fact hard to locate (Strong *et al.*, 1997). Arguably, biologists analyze poor quality data every day, which may make a plausible argument for allowing 100 percent access to all data, even the poor data, because this is important to the domain. "Big data" models may tolerate lower data quality in favor of massive increases in data quantity.

Similarly, there is a distinction between being mutually understood and logically sound data interpretation. Bioinformatics scientists from the domains of biology and computer science, care more about the interpretation of the curation data to the extent that data is recorded in appropriate languages, symbols, units, and the degree to which definitions and classifications are clear. Data and information can be mutually understandable within a user group, but may not be interpretable outside that group because of unfamiliarity of specific language, scientific symbols, and data formatting structure. Genomics scientists with knowledge of bioinformatics ranked interpretability significantly higher than did the two other groups. It is also presumed that scientists occupying both biology and computer science domains, do in fact command sufficient knowledge of both fields to "assess the integrity of the data and to grasp their meaning" (Borgman, 2012, p. 1072). Computational scientists were found to care more about understandability of the curation records than other two groups. They might focus on offering help to design user analysis tools for better use/reuse of curated data. Computational scientists usually require more insightful biological knowledge, background readings, and reference materials to ensure their data curation that makes good biological sense and is understandable, both in intermediate and final curated records/outputs.

The curated data sources could be lab reports, field notes, archival records, and other information objects. Genomics scientists have to use various sources of information to digitize and integrate the disparate pieces of genomic data. The represented curation should be concise and well-organized, as "one-point access" of a richly curated repository (Chilana *et al.*, 2009, p. 76). Computational scientists ranked Concise representation highly, and believed the improvement of the genome curation and its data representation in a concise and coherent fashion could improve understandability of data, and therefore reduce the burden of the flood of information being processed. Scientists with domain knowledge of computer science or bioinformatics could aid the development of data formats and metadata standards to support both external data linkage, and heterogeneous data referencing. Survey data suggested that the usefulness of curated data could improve the support of user-friendly browsing, retrieving and data manipulation in an online collaborative environment. Similar findings were also observed in the following open-ended survey comments:

> Moving between concise and detailed representations may be helpful.
>
> Having Graphics [is] nice [to browse].
>
> These accessions should have been linked to the page.
>
> Well described but data not structured; therefore it will be difficult to parse in automatic ways.

In addition, currency ("Up-to-date") was ranked highly by biologists. It might be concluded that curated genomic data should be frequently updated and reassessed

because of the rapid changing nature of biological knowledge (Huang *et al.*, 2012). Curated information should be the most current information, with interoperation from different database platforms, as the examples below from the survey comments show:

> Cross-compatibility with other public database, and the up-to-date relevant linkage to external databases […]
>
> This curation record has the most recent detail as both protein and CDS sequences are available with accessible hyperlinks.

Domain knowledge and DQ skills perceptions

Curating genomic data requires highly-developed interdisciplinary skills, including a capacity for critical thinking and problem solving, and for cross-disciplinary thinking. Most of the scientists in this study are scholars with PhDs (81 percent), obtaining educational training or research experience in either biology or other related fields. It also requires skills in information, communication, and technology. Biological experts have a high level of proficiency in domain knowledge – biology. They are more confident in making judgments, evaluation, or comments for curation program outputs (Chilana *et al.*, 2009). They are good at interpreting curation results, but need to consult computational scientists or programmers to obtain complicated programming tools for data mining, the switching between different database platforms, and the locating of relevant curation resources. Computational scientists offer technical support and translate curation problems into actionable programming tools. They need to work closely with biologists to ensure that their curation program outputs are consistent with the original biological problem (Chilana *et al.*, 2009). The survey data suggests a trend in genome curation work for the engagement of more scientists with both computer science and biology domain knowledge. Bioinformaticians with knowledge of both domains have advantages integrating biological knowledge into applicable solutions of curation. Computational scientists regard the use of curation of data mining and database tools (e.g. Data warehousing, SQL) as important, since their jobs involve data wrangling, integration and retrieval in large-scale databases (Heer and Kandel, 2012).

All three user groups value highly the DQ technical skills. This finding indicated that curating genomic data requires a great number of data mining and statistical analysis tools to support data curation related tasks. The DQ adaptive skills mattered more to computational scientists than the other two groups (Figure 2). Adaptive skills are those that allow computational scientists or programmers to actively interact with other users, which helps facilitate understanding of users' requirements, and translate the curation problems into practical solutions.

Data quality literacy reflects the ability of users to understand data quality related concepts, knowledge and skills. Particularly, data quality literacy skills such as DQ dimensions and DQ implication were ranked highly among biologists. This finding suggests that grasping the necessary knowledge of data quality concepts, the related assessment methods, and their ways to identify potential data quality problems are prerequisites for scientists to secure high quality curation work. Computational scientists and biologists might have sufficient skills when operating with their own domains, but they might be also interested in the cross-disciplinary skills required for scientific data management and data quality assurance. Such skills, as well as other annotation and data mining skills, could facilitate the curation activities, data quality assurance, and data provenance services in genome curation work.

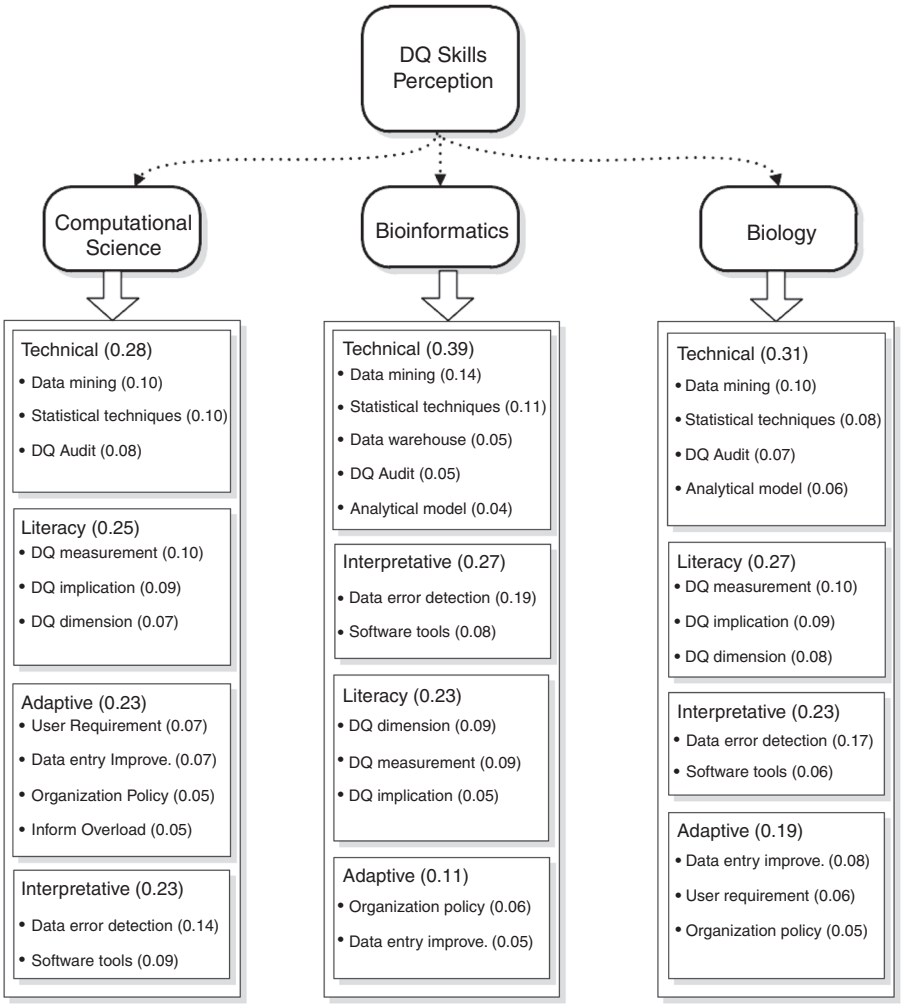


Figure 2.
Domain knowledge based DQ skills grouping priorities

Note: Only those with cumulative rankings of less than 90 percent were kept (see Appendix 3)

The trends regarding quality assurance and knowledge creation activities (e.g. data annotation) ultimately evolves into higher expectations for bioinformatics literacy, including data quality literacy on the part of users. According to a recent survey (Burge *et al.*, 2012), the biocuration community believes that a genome curator, having both research experience and a strong biological or computational background, would benefit their work tremendously. The differences among users' perceptions need to be benchmarked, collected, and communicated. The empirically-based community feedback is needed to design appropriate strategies for improvement in curation quality.

The findings of this research can help develop curation domain-specific data quality models. Computational scientists ranked usefulness higher than accessibility, whereas biologists did the opposite (Figure 1). This may also suggest the presence of trade-offs

among different data quality dimensions is related to the users' domain knowledge. Literature suggests that data quality activities are not free, it requires the user's priorities, including participation in a possible trade-offs based on the different dimensions of quality optimization (Ballou and Pazer, 1995; Stvilia *et al.*, 2007). The identified data quality trade-offs are reasonable. This is provided that the overall data quality is of sufficient and good enough for its use in the research practice. For example, accessibility is expensive when dealing with a high volume of the data. These data quality trade-offs can be justified when organized and curated data is difficult to obtain and access with given time restrictions. Biologists required sufficient computational skills or knowledge to access and retrieve the data they want, but they may have to accept and tolerate raw and unstructured curated data in exchange for having timely access to important information. Similarly, computational scientists possess advanced skills in genome related data wrangling (Heer and Kandel, 2012), they focus on making their collected data more usable by adding more curated information.

Genomics scientists, like scholars from other scientific disciplines, require sufficient data curation and process skills to conduct tremendous data manipulation work. This study collected empirical data through a survey of members in a particular scientific community. It reports members' perceived priorities for data quality criteria and identified related DQ skills in the context of genomic data curation work (see Figures 1 and 2). The findings of this study can be used in the development of genomic data curation procedures, policies, and training modules. These curation artifacts could be used by the current curation team and by future institutional end-users and participants, who may themselves not possess extensive trainings in data curation and data management.

Conclusion

The way scientists solve problems in genome curation today is probably not the way scholars and practitioners did so a decade ago. Since technology is growing, our knowledge and abilities are also increasing, and our analytical methods are changing as well. Genome curation work is a collaborative process executed through a dynamic complex interaction among those scientists who hold diverse domain knowledge and work experience. It requires scientists to read tremendous amounts of research literature, and to obtain solid domain knowledge. It also requires scientists to be flexible and adaptive to deal with different scales of genomic related data, to make sound judgments regarding the annotated information in the genome context, and to ensure the capture of all related information within the data model.

Scientists' domain knowledge and experience in genome curation work eventually impacts their priorities for the data quality criteria. Overall, scientists must process enormous amounts of distributed data through many different tools developed to aid them in knowledge discovery. This work will allow for richer knowledge representation and manipulation. This study also has some limitations. The data was collected by survey, rather than direct observation to collect the opinions of the scientists regarding data quality skills and dimensions requirements used to develop the data quality models for genome curation. The data are therefore only approximations of the respondents' actual value models for quality and for data quality skills used in practice. Future research collection of additional empirical data through observations and interviews can help determine the community's data curation and quality assurance practices. What is more, the importance of these concepts was recorded by survey participants at the time of survey completion; the follow-up interview provides an

opportunity to validate where modifications are necessary. It should also be pointed out that the data quality skills used as the survey instrument were in fact based on previous studies (Chung *et al.*, 2002). As new data management technologies evolve (e.g. computing with graphics processing units and "cloud" technologies), these items and related constructs may require a revisit to update the priorities of the community regarding data quality assurances skills.

Genomics research is data-intensive. Some significant differences were observed in scientists' perception of data quality requirements in genome curation work which required calibration of their knowledge across different domains. This study found that given a common curation task with the same data-quality information, genomics researchers with diverse domain knowledge make different decisions regarding data-quality trade-offs. Through this study, the identification of the variations of the DQ models based on domain knowledge can help better understand the function of data quality in context of domain knowledge. It can also help identify related curation tools and supports for the genomics research community, and to develop curation policies, procedures, training modules, and strategies, and problem-solving paths tailored to the curation work. Future studies could involve the collection of additional data and the development of operational models of these trade-offs, allowing them being used in practice to optimize quality assurance activities.

References

Allen, B. (1991), "Topic knowledge and online catalog search formulation", *Library Quarterly*, Vol. 61 No. 2, pp. 188-213.

Alexander, P.A. (1992), "Domain knowledge: evolving themes and emerging concerns", *Educational Psychologist*, Vol. 27 No.1, pp. 33-51.

Ballou, D.P., and Pazer, H.L. (1995), "Designing information systems to optimize the accuracy-timeliness tradeoff", *Information Systems Research*,Vol. 6 No.1, pp. 51-72.

Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch-Mizrachi, I. and Ostell, J. (2012), "Bioproject and BioSample databases at NCBI: facilitating capture and organization of metadata", *Nucleic acids research*, Vol. 40 No. D1, pp. D57-D63.

Bartlett, J.C. and Toms, E.G. (2005), "Developing a protocol for bioinformatics analysis: an integrated information behavior and task analysis approach", *Journal of the American Society for Information Science and Technology*, Vol. 56 No. 5, pp. 469-482.

Bentley, D.R. (2006), "Whole-genome re-sequencing", *Current Opinion in Genetics & Development*, Vol. 16 No.6, pp. 545-552.

Borgman, C.L. (2012). "The conundrum of sharing research data", *Journal of the American Society for Information Science and Technology*, Vol. 63 No. 6, pp. 1059-1078.

Brent, M.R. (2008), "Steady progress and recent breakthroughs in the accuracy of automated genome annotation", *Nature Reviews Genetics*, Vol. 9 No. 1, pp. 62-73.

Brown, C.M. (2003), "The changing face of scientific discourse: analysis of genomic and proteomic database usage and acceptance", *Journal of the American Society for Information Science & Technology*, Vol. 54 No. 10, pp. 926-938.

Burge, S., Attwood, T.K., Bateman, A., Bateman, A., Berardini, T.Z., Cherry, M. and Gaudet, P. (2012), "Biocurators and biocuration: surveying the 21st century challenges", *Database*, bar059, doi:10.1093/database/bar059.

Burkhardt, K., Schneider, B. and Ory, J. (2006), "A biocurator perspective: annotation at the research collaboratory for structural bioinformatics protein data bank", *PLoS Computational Biology*, Vol. 2 No. 10, p. e99.

Carroll, J.M. (1997), "Scenario-based design", in Helander, M. and Landauer, T.K. (Eds), *Handbook of Human–Computer Interaction*, North Holland, Amsterdam, pp. 383-406.

Chilana, P.K., Palmer, C.L. and Ko, A.J. (2009), "Comparing bioinformatics software development by computer scientists and biologists: an exploratory study", *Software Engineering for Computational Science and Engineering, SECSE '09. ICSE Workshop*, Vol. 72 No. 79, pp. 23-24.

Chung, W., Fisher, C. and Wang, R. (2002), "What skills matter in data quality?", *The 7th International Conference on Information Quality (ICIQ-'02), Boston, MA*.

Cole, N.J. and Bawden, D. (1996), "Bioinformatics in the pharmaceutical industry", *Journal of Documentation*, Vol. 52 No. 1, pp. 51-68.

Diaper, D. (2004), "Understanding task analysis in human computer interaction", in Diaper, D. and Stanton, N. (Eds), *The Handbook of Task Analysis for Human–Computer Interaction*, Erlbaum, Mahwah, NJ, pp. 117-133.

Dukerich, J.M. and Nichols, M.L. (1991), "Causal information search in managerial decision making", *Organizational Behavior and Human Decision Processes*, Vol. 50 No. 1, pp. 106-122.

Evans, J.R. and Lindsay, W.M. (2005), *The Management and Control of Quality*, Thomson Learning, Cincinnati, OH, pp. 132-136.

Eysenck, H.J. (1991), "Dimensions of personality: 16, 5 or 3?—criteria for a taxonomic paradigm", *Personality and Individual Differences*, Vol. 12 No. 8, pp. 773-790.

Fenstermacher, D. (2005), "Introduction to bioinformatics", *Journal of the American Society for Information Science and Technology*, Vol. 56 No. 5, pp. 440-446.

Fisher, C.W., Chengalur-Smith, I. and Ballou, D.P. (2003), "The impact of experience and time on the use of data quality information in decision making", *Information Systems Research*, Vol. 14 No. 2, pp. 170-188.

Frické, M. and Fallis, D. (2004), "Indicators of accuracy for answers to ready reference questions on the internet", *Journal of the American Society for Information Science and Technology*, Vol. 55 No. 3, pp. 238-245.

Gilliland, S.W., Wood, L. and Schmitt, N. (1994), "The effects of alternative labels on decision behavior: the case of corporate site selection decisions", *Organizational Behavior and Human Decision Processes*, Vol. 58 No. 3, pp. 406-427.

Go, K. and Carroll, J. (2004), "Scenario-based task analysis", in Diaper, D. and Stanton, N. (Eds), *The Handbook of Task Analysis for Human-Computer Interaction*, Erlbaum, Mahwah, NJ, pp. 117-133.

Goth, G. (2012), "Preserving digital data", *Communications of the ACM*, Vol. 55 No. 4, pp. 11-13.

Haythornthwaite, C. (2006), "Learning and knowledge networks in interdisciplinary collaborations", *Journal of the American Society for Information Science and Technology*, Vol. 57 No. 8, pp. 1079-1092.

Heer, J. and Kandel, S. (2012), "Interactive analysis of big data", *XRDS: Crossroads, The ACM Magazine for Students*, Vol. 19 No. 1, pp. 50-54.

Hemminger, B.M., Saelim, B., Sullivan, P.F. and Vision, T.J. (2007), "Comparison of full-text searching to metadata searching for genes in two biomedical literature cohorts", *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 14, pp. 2341-2352.

Hjørland, B. and Albrechtsen, H. (1995), "Toward a new horizon in information science: domain-analysis", *Journal of the American Society for Information Science and Technology*, Vol. 46 No. 6, pp. 400-425.

Huang, H., Andrews, J. and Tang, J. (2012), "Citation characterization and impact normalization in bioinformatics journals", *Journal of the American Society of Information Science and Technology*, Vol. 63 No. 3, pp. 490-497.

Huang, H., Lu, J., Hunter, W. and Liang, S. (2011), "Using IBM Content Manager for genomic data annotation and quality assurance tasks", *IBM Journal of Research and Development*, Vol. 55 No. 6, p. 13.

Huang, H., Stvilia, B., Jörgensen, C. and Bass, H. (2012), "Prioritization of data quality dimensions and skills requirements in genome annotation work", *Journal of the American Society for Information Science and Technology*, Vol. 63 No. 1, pp. 195-207.

Ioannidis, J.P. and Khoury, M.J. (2011), "Improving validation practices in 'omics' research", *Science*, Vol. 334 No. 6060, pp. 1230-1232.

Ju, B. (2007), "Does domain knowledge matter: mapping users' expertise to their information interactions", *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 13, pp. 2007-2020.

Klein, B.D., Goodhue, D.L. and Davis, G.B. (1997), "Can humans detect errors in data? Impact of base rates, incentives, and goals", *MIS Quarterly*, Vol. 21 No. 2, pp.169-194.

Kuhn, T.S. (1974), "Second thoughts paradigms", in Suppe, F. (Ed.), *The Structure of Science Theories*, University of Illinois Press, London, pp. 459-482.

Lankes, R.D. (2008), "Credibility on the internet: shifting from authority to reliability", *Journal of Documentation*, Vol. 64, No. 5, pp. 667-686.

Lathe, W., Williams, J., Mangan, M. and Karolchik, D. (2008), "Genomic data resources: challenges and promises", *Nature Education*, Vol. 13 No. 3, p. 2.

Lee, Y. and Strong, D. (2003), "Knowing – why about data processes and data quality", *Journal of Management Information Systems*, Vol. 20 No. 3, pp. 13-39.

Leonelli, S., Diehl, A.D., Christie, K.R., Harris, M.A. and Lomax, J. (2011), "How the gene ontology evolves", *BMC Bioinformatics*, Vol. 12, No. 1, p. 325.

Mackay, J.M. and Elam, J.J. (1992), "A comparative study of how experts and novices use a decision aid to solve problems in complex knowledge domains", *Information Systems Research*, Vol. 3 No. 2, pp. 150-172.

MacMullen, W. (2006), "Contextual analysis of variation and quality in human-curated gene ontology annotations", PhD dissertation, University of North Carolina.

MacMullen, W.J. and Denn, S.O. (2005), "Information problems in molecular biology and bioinformatics", *Journal of the American Society for Information Science and Technology*, Vol. 56 No. 5, pp. 447-456.

Marchionini, G., Dwiggins, S., Katz, A. and Lin, X. (1993), "Information seeking in full-text end-user-oriented search systems: the roles of domain and search expertise", *Library & Information Science Research*, Vol. 15, No. 1, pp. 35-69.

Mayor, C. and Robinson, L. (2014), "Ontological realism, concepts and classification in molecular biology: development and application of the gene ontology", *Journal of Documentation*, Vol. 70 No. 1, pp. 173-193.

Owrang, O.M.M. and Grupe, F.H. (1996), "Using domain knowledge to guide database knowledge discovery", *Expert Systems With Applications*, Vol. 10 No. 2, pp. 173-180.

Pagani, I., Liolios, K., Jansson, J., Chen, I.M.A., Smirnova, T., Nosrat, B. and Kyrpides, N.C. (2012), "The genomes online database (GOLD) v. 4: status of genomic and metagenomic projects and their associated metadata", *Nucleic Acids Research*, Vol. 40 No. D1, pp. D571-D579.

Palmer, C.L. and Neumann, L.J. (2002), "The information work of interdisciplinary humanities scholars: exploration and translation", *The Library Quarterly*, Vol. 72 No. 1, pp. 85-117.

Payne, J.W., Bettman, J.R. and Johnson, E.J. (1993), *The Adaptive Decision Maker*, University Press, Cambridge.

Pruitt, K.D., Tatusova, T., Brown, G.R. and Maglott, D.R. (2012), "NCBI reference sequences (RefSeq): current status, new features and genome annotation policy", *Nucleic Acids Research*, Vol. 40 No. D1, pp. D130-D135.

Reed, J.L., Famili, I., Thiele, I. and Palsson, B.O. (2006), "Towards multidimensional genome annotation", *Nature Reviews Genetics*, Vol. 7 No. 2, pp. 130-141.

Rieh, S. (2002), "Judgment of information quality and cognitive authority in the web", *Journal of the American Society for Information Science and Technology*, Vol. 53 No. 2, pp. 145-161.

Rouet, J.-F., Favart, M., Britt, M.A. and Perfetti, C.A. (1997), "Studying and using multiple documents in history: effects of discipline expertise", *Cognition and Instruction*, Vol. 15, No. 1, pp. 85-106.

Salimi, N. and Vita, R. (2006), "The biocurator: connecting and enhancing scientific data", *PLoS Computational Biology*, Vol. 2 No. e125, p. 2, available at: www.nature.com/scitable/topicpage/genomic-data-resources-challenges-and-promises-743721

Samuel, V., Gussman, A. and Klumke, W. (2008), "Toward an online repository of standard operating procedures (SOPs) for (meta)genomic annotation", *OMICS: A Journal of Integrative Biology*, Vol. 12 No. 2, pp. 137-141.

Sanbonmatsu, D.M., Kardes, F.R. and Herr, P.M. (1992), "The role of prior knowledge and missing information in multiattribute evaluation", *Organizational Behavior and Human Decision Processes*, Vol. 51 No. 1, pp. 76-91.

Sanderson, K. (2011), "Bioinformatics: curation generation", *Nature*, Vol. 470 No. 7333, pp. 295-296.

Shachak, A. and Fine, S. (2008), "The effect of training on biologists acceptance of bioinformatics tools: a field experiment", *Journal of the American Society for Information Science and Technology*, Vol. 59 No. 5, pp. 719-730.

Shimoyama, M., Hayman, G.T., Laulederkind, S.J.F., Nigam, R., Lowry, T.F. and Dwinell, M.R. (2009), "The rat genome database curators: who, what, where, why", *PLoS Computational Biology*, Vol. 5 No. e1000582, p. 2, available at: www.nature.com/scitable/topicpage/genomic-data-resources-challenges-and-promises-743721

Strong, D., Lee, Y. and Wang, R. (1997), "Data quality in context", *Communication of the ACM*, Vol. 40 No. 5, pp. 103-110.

Stvilia, B., Gasser, L., Twidale, M., and Smith, L. (2007), "A framework for information quality assessment", *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 12, pp. 1720-1733.

Stvilia, B., Twidale, M., Smith, L.C. and Gasser, L. (2008), "Information quality work organization in Wikipedia", *Journal of the American Society for Information Science and Technology*, Vol. 59 No. 6, pp. 983-1001.

Tabatabai, D. and Shore, B.M. (2005), "How experts and novices search the web", *Library & Information Science Research*, Vol. 27 No. 2, pp. 222-248.

Vibert, N., Ros, C., Bigot, L.L., Ramond, M., Gatefin, J. and Rouet, J.F. (2009), "Effects of domain knowledge on reference search with the PubMed database: an experimental study", *Journal of the American Society for Information Science and Technology*, Vol. 60 No. 7, pp. 1423-1447.

Vibert, N., Rouet, J.-F., Ros, C., Ramond, M. and Deshoullières, B. (2007), "The use of online electronic information resources in scientific research: the case of neuroscience", *Library & Information Science Research*, Vol. 29 No. 4, pp. 508-532.

Wang, R. and Strong, D. (1996), "Beyond accuracy: what data quality means to data consumers", *Journal of Management Information Systems*, Vol. 12 No. 4, pp. 5-35.

Watt, W.B. (2000), "Avoiding paradigm-based limits to knowledge of evolution", in Clegg, M.T., Hecht, M.K. and MacIntyre, R.J. (Eds), *Evolutionary Biology*, Springer, pp. 73-96.

Wildemuth, B.M. (2004), "The effects of domain knowledge on search tactic formulation", *Journal of the American Society for Information Science and Technology*, Vol. 55 No. 3, pp. 246-258.

Willis, C., Greenberg, J. and White, H. (2012), "Analysis and synthesis of metadata goals for scientific data", *Journal of the American Society for Information Science and Technology*, Vol. 63 No. 8, pp. 1505-1520.

Wooley, J.C. and Lin, H.S. (Eds) (2005), *Catalyzing Inquiry at the Interface of Computing and Biology*, National Academies Press, Washington, DC.

Wu, J., Zhang, Y., Zhang, H., Huang, H., Folta, K. and Lu, J. (2010), "Whole genome wide expression profiles of *vitis amurensis* grape responding to downy mildew by using Solexa sequencing technology", *BMC Plant Biology*, Vol. 10 No. 1, p. 234.

Wu, L.L., Huang, M.H. and Chen, C.Y. (2012), "Citation patterns of the pre-web and web-prevalent environments: the moderating effects of domain knowledge", *Journal of the American Society for Information Science and Technology*, Vol. 63 No. 11, pp. 2182-2194.

Yang, X., Ye, Y., Wang, G., Huang, H., Yu, D. and Liang, S. (2011), "VeryGene: linking tissue-specific genes to diseases, drugs, and beyond for knowledge discovery", *Physiological genomics*, Vol. 43 No. 8, pp. 457-460.

Yates, J.F., McDaniel, L. S. and Brown, E.S. (1991), "Probabilistic forecasts of stock prices and earnings: the hazards of nascent expertise", *Organizational Behavior and Human Decision Processes*, Vol. 49 No. 1, pp. 60-79.

Further reading

Mao, J.Y. and Benbasat, I. (2000), "The use of explanations in knowledge-based systems: cognitive perspectives and a process-tracing analysis", *Journal of Management Information Systems*, Vol. 17 No. 2, pp. 153-180.

Salzberg, S. (2007), "Genome re-annotation: a wiki solution?", *Genome Biology*, Vol. 8 No. 1, p. 102.

Appendix 1. Two genome curation scenarios

Scenario 1: production, curation, and submission of expressed sequence tags (ESTs) data

In this scenario, you will generate primary sequence data. For this purpose, you will process, curate, annotate, and submit sequence data as annotated sequence records in a public database. Specifically, you will produce a cDNA library, and obtain 1,000 random sequence reads (ESTs) from that cDNA library. The library contains clones from a model organism for which a genome sequence is publicly available. As part of preparing these annotated records, you will be taking steps which include annotation and data quality assurance steps to:

- process the raw data to remove vector or low quality sequences;
- annotate the sequences with regards to the genome location;
- predict gene products using routine bioinformatic tools such as BLAST alignments, open reading frames (ORFs) predictions, and comparison of predicted proteins to protein motif databases;
- produce additional annotation to link these predicted gene products to gene ontology, molecular networks, or biochemical pathways; and
- submit these ESTs and associated annotations to two different databases, GenBank, and your species specific database.

*The phrase "sequence records" refers to both the primary DNA sequences themselves and all the associated annotations.

Scenario 2: whole genome data curation in a model organism

In this scenario, you will generate genome annotation records for a particular model organism. You will use the full spectrum of genome annotation approaches including: predicted gene and protein annotation, sequences comparisons and alignments, genome variations analysis, the organization and annotation of molecular networks and biochemical pathways. You will employ these approaches using specialized databases, bioinformatics software, and literature mining to:

(1) Create sequence records for release to the public:

- Curate, annotate genome sequence data features from the sequence data by identifying the gene features (e.g. promoters, gene length, terminators) and genomic properties (e.g. motifs, repeats) from the sequence data.
- Create explicit comments to the sequence data organized along a schema that needs to be specified (e.g. gene name, gene function, enzyme identifier, bibliographic reference, experimentally identified feature, ESTs, etc.)
- Compare, correct, reannotate, or externally link the sequence data to the data available in other databases or scientific literature.

(2) Conduct data quality control by corresponding with collaborators regarding missing or inaccurate information.

(3) Assist in problem identification and recommend enhancements to the procedures in genome annotation work.

*These two scenarios were adopted from Huang *et al.* (2012).

Appendix 2

Table AI.
Cumulated percentiles of the ranked DQ dimensions in domain knowledge

	Bioinfo			Biology			Both	
	%	Cumulated %		%	Cumulated %		%	Cumulated %
Accessibility	14.7	14.7	Accuracy	15.5	15.5	Accuracy	14.9	14.9
Accuracy	14.7	29.3	Accessibility	14	29.5	Accessibility	11.7	26.6
Completeness	8.6	37.9	Completeness	10.4	40	Completeness	10.6	37.2
Understandability	8.6	46.6	Believability	8.2	48.2	Believability	9.0	46.3
Appro amount of info	6.0	52.6	Up-to-date	8.0	56.2	Interpretability	8.5	54.8
Believability	6.0	58.6	Consistency	7.3	63.4	Consistency	7.4	62.2
Consistency	6.0	64.7	Appro amount of info	6.3	69.7	Appro amount of info	5.3	67.6
Interpretability	6.0	70.7	Ease of manipulate	5.6	75.3	Unbiased	5.3	72.9
Unbiased	6.0	76.7	Traceability	5.6	80.9	Ease of manipulate	4.8	77.7
Up-to-date	6.0	82.8	Unbiased	5.1	86	Up-to-date	4.8	82.4
Concise repres	4.3	87.1	Interpretability	3.4	89.3	Traceability	4.3	86.7
Ease of manipulate	4.3	91.4	Understandability	3.1	92.5	Understandability	3.7	90.4
Relevance[a]	*2.6*	*94*	*Relevance*	*1.7*	*94.2*	*Value-added*	*2.7*	*93.1*
Traceability	*2.6*	*96.6*	*Security*	*1.7*	*95.9*	*Relevance*	*2.1*	*95.2*
Reputation	*1.7*	*98.3*	*Value-added*	*1.7*	*97.6*	*Reputation*	*2.1*	*97.3*
Value-added	*1.7*	*100*	*Reputation*	*1.5*	*99*	*Concise repres*	*1.6*	*98.9*
Security	*0.0*	*100*	*Concise repres*	*1.0*	*100*	*Security*	*1.1*	*100*

Note: [a]DQ dimensions are italics indicating their accumulated rankings over 90 percent

Appendix 3

	Bioinfo			Biology			Both	
	%	Cumulated %		%	Cumulated %		%	Cumulated %
Data-error detection	13.6	13.6	Data-error detection	15	15	Data-error detection	17.5	17.5
DQ measurement	9.1	22.7	DQ measurement	8.8	23.8	Data-mining skills	13.5	31
Statistical techniques	9.1	31.8	Data-mining skills	8.8	32.6	Statistical techniques	9.9	40.9
Data-mining skills	9.1	40.9	Statistical techniques	7.7	40.3	DQ dimensions	8.8	49.7
DQ implication	8.2	49.1	DQ implication	7.8	48.1	DQ measurement	8.2	57.9
DQ audit	8.2	57.3	DQ dimensions	7.5	55.6	Software tools	7.0	64.9
Software tools	8.2	65.5	Data-entry improvement	6.9	62.5	Organization policies	5.9	70.8
DQ dimensions	6.4	71.8	DQ audit	6.2	68.7	Data-warehouse set-up	5.2	76
Data-entry improvement	6.4	78.2	Software tools	6.0	74.7	DQ implication	4.7	80.7
User requirement	6.4	84.6	User requirement	5.9	80.6	Data-entry improvement	4.1	84.8
Organization policies	4.5	89.1	Analytic models	4.9	85.5	DQ audit	4.1	88.9
Information overload	4.5	93.6	Organization policies	4.2	89.7	Analytic models	3.5	92.4
Analytic models[a]	*2.8*	*96.4*	*Data-warehouse set-up*	*4.1*	*93.8*	*User requirement*	*2.9*	*95.3*
Change process	*1.8*	*98.2*	*Change process*	*2.8*	*96.6*	*DQ cost/benefit*	*1.8*	*97.1*
SQL	*1.8*	*100*	*Information overload*	*1.9*	*98.5*	*Change process*	*1.1*	*98.2*
DQ cost/benefit	*0.0*	*100*	*DQ cost/benefit*	*1.5*	*100*	*Information overload*	*1.2*	*99.4*
Data-warehouse set-up	*0.0*	*100*	*SQL*	*0.0*	*100*	*SQL*	*0.6*	*100*

Note: [a]DQ skills are italics indicating their accumulated rankings over 90 percent

Table AI.
Cumulated percentiles of the ranked DQ skills in domain knowledge

About the author
Dr Hong Huang is an Assistant Professor at the School of Information at the University of South Florida. He received MS Degrees in Genetics and Computer Science, and PhD in Library and Information Studies from the Florida State University. His primary research interests included scientist behaviors, scientific data management, and bioinformatics. He also conducts research related to scholarly communication, and health information behaviors. He is currently interested in understanding the behaviors of library users and their interactions with social media tools. Dr Hong Huang can be contacted at: honghuang@usf.edu

Ann Surg Oncol (2013) 20:1759–1765
DOI 10.1245/s10434-013-2901-1

Annals of
SURGICAL ONCOLOGY
OFFICIAL JOURNAL OF THE SOCIETY OF SURGICAL ONCOLOGY

ORIGINAL ARTICLE – HEALTHCARE POLICY AND OUTCOMES

Comparison of Cases Captured in the National Cancer Data Base with Those in Population-based Central Cancer Registries

Catherine C. Lerro, MPH[1], Anthony S. Robbins, MD, PhD, MPH[2], Jerri Linn Phillips, MA[3], and Andrew K. Stewart, MA[3]

[1]Department of Environmental Health Sciences, Yale University School of Public Health, New Haven, CT; [2]Health Services Research Program, American Cancer Society, Atlanta, GA; [3]Commission on Cancer, American College of Surgeons, Chicago, IL

ABSTRACT

Background. The National Cancer Data Base (NCDB) is a large, geographically diverse hospital-based cancer registry that has been used to study factors related to cancer diagnosis, treatment, and survival. The primary purpose of this study was to compare the case counts and characteristics of patients in NCDB with population-based registries reported in the United States Cancer Statistics (USCS).

Methods. Cancer case counts from NCDB were compared to case counts from USCS to measure NCDB's case coverage, or the percentage of cases captured. Case coverage was examined by a variety of characteristics, including state of residence, race/ethnicity, age, and primary cancer site.

Results. The overall NCDB case coverage was 67.4 %, ranging from a high of 88.7 % for Delaware to a low of 27.1 % for Arizona. Case coverage for white, black, and Asian/Pacific Islander cases was high (64.7 % to 67.4 %), but it was much lower for American Indians/Alaskan Natives (32.8 %) and those of Hispanic ethnicity (51.1 %). Among the elderly (aged 65 + years), case coverage is much lower compared to persons younger than 65 (63.0 % and 73.0 %, respectively). Case coverage also varied widely by site, with the highest being cervix (77.9 %) and the lowest being melanoma (50.6 %).

Conclusions. This study highlights the geographic- and site-specific variation in NCDB case coverage, primarily as a result of NCDB facility presence and data collection and processing protocols. These findings illustrate the strengths and limitations of NCDB as a resource for nationwide data on cancer diagnosis, treatment, and survival.

First Received: 14 June 2012;
Published Online: 9 March 2013

A. S. Robbins, MD, PhD, MPH
e-mail: anthony.robbins@cancer.org

The National Cancer Data Base (NCDB) is a hospital-based cancer registry that is jointly sponsored by the American College of Surgeons' Commission on Cancer (CoC) and the American Cancer Society (ACS).[1] NCDB collects data on newly diagnosed cancer patients diagnosed and treated at over 1500 CoC-accredited facilities throughout the United States, and earlier studies estimate that it captures 70 % of all cancer patients diagnosed annually in the United States.[1,2]

Investigators at ACS and the American College of Surgeons have conducted many studies using NCDB, examining associations between stage at diagnosis, treatment, and survival by health insurance status and race/ethnicity. An important strength of these studies is the inclusion of a large and geographically diverse sample of cancer patients. There are CoC-accredited facilities in every state, the District of Columbia, and Puerto Rico; Wyoming was the final state to add a CoC-approved facility in 2010. The CoC accreditation program encourages facilities to improve their quality of patient care through various cancer-related programs focusing on prevention, early diagnosis, pretreatment evaluation, staging, optimal treatment, rehabilitation, surveillance for recurrent disease, support services, and end-of-life care.[3] Patients treated at these facilities and reported to NCDB reside in all 50 states, the District of Columbia, and Puerto Rico. However, one important limitation has been a lack of systematic data assessing the coverage of this sample with respect to all cancer cases in the United States. Bilimoria et al. compared NCDB case counts by cancer site to estimates produced for ACS's 2005 *Cancer Facts and Figures*.[2,4] These analyses suggested substantial variation

in NCDB coverage across cancer sites, but the comparison of hospital registry data to projected estimates of incident cases may not be ideal, particularly for low-incidence cancer sites.[4] Moreover, the authors of the *Cancer Facts and Figures* have since made significant alterations to improve the methodology for estimating cancer incidence counts.[5] The prior methodology did not allow for geographic variation in incidence rates; instead, it extrapolated incidence rates for the entire country using data from the Surveillance Epidemiology and End Results (SEER) 9 registries, which represent only approximately 10 % of the population. Bilimoria et al. also examined CoC-approved facilities that use American Hospital Association Annual Survey data and found wide variation in geographic coverage by state.[6]

The primary purpose of this study was to compare case counts and characteristics of NCDB cases with those reported in the United States Cancer Statistics (USCS).[7] USCS is a compilation of incidence data from registries participating in the National Program of Cancer Registries (NPCR) and SEER. USCS case counts are considered to be the gold standard because they are obtained from population-based registries.

METHODS

Analyses included cases diagnosed in 2004 to 2006; all 50 states and the District of Columbia met USCS publication criteria for these years.[8,9] USCS data are publicly available.[10] Analyses included all primary malignancies (excluding nonmelanoma skin), benign brain and central nervous system tumors, and in situ urinary bladder tumors.

Cancer case counts from NCDB were compared to case counts from USCS in order to measure NCDB's case coverage, or the number of cases in NCDB divided by number of cases in USCS. The NCDB is a hospital-based registry comprising CoC-accredited hospitals that are required to report annually all analytic cases to the NCDB.[11] The NCDB contains data elements on patient demographics, insurance status, tumor characteristics, first course of treatment, ZIP code–level socioeconomic factors, and facility-level characteristics. Data reporting to NCDB is highly standardized and similar to other federal cancer registry data systems, such as SEER and NPCR.[2] Case coverage was examined by state of residence, race/ethnicity, age, sex, and primary cancer site.

A metric called facility coverage, or number of CoC-approved hospitals divided by the number of total hospitals, was established in order to understand the proportional presence of NCDB facilities in a given state. The number of CoC-approved hospitals was determined from the NCDB file for all cases diagnosed in 2004 to 2006 outlined above; distinct facility identifiers were counted for each state. The Area Resource File, a publicly available data source that contains variables relating to health facilities, was used to determine the total number of facilities in each state for 2006, the most recent year examined in this study.[12] Only short-term medical and surgical facilities were included. Long-term facilities, psychiatric institutions, children's hospitals, specialty hospitals, and institutes for mental retardation were excluded.[6]

NCDB cases were restricted to match the characteristics of the USCS data. NCDB cases in patients residing in Puerto Rico or with unspecified state of residence were excluded. The primary site of the reported cancers was categorized using the SEER Site Recode (ICD-O-3).[3,13] USCS case counts by state, age group, race, and cancer site were generated using the USCS web-based reporting system.[14] NCDB case counts and facility counts were generated using SAS software (version 9.2; SAS Statistical Institute, Cary, NC). The Morehouse University institutional review board reviewed ACS studies that used the NCDB and determined that this study was exempt from review.

RESULTS

The overall NCDB case coverage was 67.4 %, ranging from a high of 88.7 % for Delaware to a low of 27.1 % for Arizona. Regional variation is apparent as well (Table 1). The case coverage was highest in the Northeast at 70.2 %, and lowest in the West at 55.9 % (results not shown). In general, states in which a higher proportion of short-term medical and surgical facilities are CoC accredited had a corresponding higher percentage of cancer patients reported to NCDB (Fig. 1). Some exceptions to this observation were noted; for example, North Dakota had 88.5 % case coverage (the second highest of any state) but 13.6 % facility coverage.

Only small differences in case coverage were observed among white, black, and Asian/Pacific Islander cases. Noticeably, NCDB case coverage is low among American Indians/Alaskan Natives (32.8 %) and Hispanics (50.1 %) compared to whites, blacks, and Asian/Pacific Islanders. NCDB case coverage was lower in the elderly, and among adults case coverage tended to decline with age (Table 2).

There was noticeable variation in NCDB case coverage by primary site (Table 3), with some sites (e.g., melanoma, Kaposi sarcoma, leukemia) having case coverage near 50 %, while for other sites (e.g., cervix, uterus, brain, nervous system), case coverage was close to 80 %. Case coverage tended to be similar for male and female subjects on a site-by-site basis; however, overall, women had a higher overall case coverage (69.3 % vs. 62.7 %;

TABLE 1 USCS and NCDB case counts, case coverage, facility counts, and facility coverage by state, 2004–2006

State	NCDB cases	USCS cases	Case coverage, %	CoC facilities	General/surgical facilities[a]	Facility coverage, %
Alabama	44,366	67,795	65.4	23	101	22.8
Alaska	4,144	7,089	58.5	3	23	13
Arizona	19,762	73,049	27.1	6	67	9.0
Arkansas	11,710	42,364	27.6	9	78	11.5
California	232,036	436,708	53.1	109	347	31.4
Colorado	40,737	55,899	72.9	21	72	29.2
Connecticut	50,336	58,608	85.9	25	30	83.3
Delaware	12,566	14,168	88.7	6	6	100.0
District of Columbia	6,823	8,133	83.9	7	9	77.8
Florida	196,635	301,806	65.2	76	197	38.6
Georgia	80,705	111,541	72.4	40	146	27.4
Hawaii	11,015	18,147	60.7	8	22	36.4
Idaho	15,005	19,129	78.4	7	38	18.4
Illinois	152,631	185,235	82.4	85	188	45.2
Indiana	70,361	90,653	77.6	48	105	45.7
Iowa	32,094	48,644	66.0	14	119	11.8
Kansas	21,414	40,803	52.5	14	130	10.8
Kentucky	52,435	68,014	77.1	26	103	25.2
Louisiana	40,360	64,283	62.8	28	109	25.7
Maine	19,717	24,936	79.1	12	37	32.4
Maryland	65,351	76,378	85.6	35	49	71.4
Massachusetts	83,880	107,961	77.7	48	65	73.9
Michigan	114,271	156,713	72.9	38	135	28.2
Minnesota	54,902	74,361	73.8	19	128	14.8
Mississippi	27,084	40,620	66.7	11	96	11.5
Missouri	65,330	88,158	74.1	36	118	30.5
Montana	9,141	14,553	62.8	6	55	10.9
Nebraska	19,194	26,691	71.9	11	84	13.1
Nevada	13,434	32,113	41.8	5	31	16.1
New Hampshire	13,798	21,049	65.6	10	26	38.5
New Jersey	113,025	140,977	80.2	56	72	77.8
New Mexico	13,317	24,229	55.0	6	43	14.0
New York	177,703	303,442	58.6	71	197	36.0
North Carolina	99,591	126,425	78.8	39	117	33.3
North Dakota	8,918	10,079	88.5	6	44	13.6
Ohio	145,527	175,910	82.7	86	158	54.4
Oklahoma	34,310	54,160	63.3	17	112	15.2
Oregon	39,865	55,647	71.6	21	59	35.6
Pennsylvania	155,490	220,094	70.6	85	166	51.2
Rhode Island	15,940	18,349	86.9	11	11	100.0
South Carolina	43,939	64,361	68.3	17	63	27.0
South Dakota	8,432	11,527	73.1	7	57	12.3
Tennessee	65,799	87,735	75.0	35	119	29.4
Texas	141,081	269,104	52.4	78	373	20.9
Utah	14,501	23,630	61.4	6	40	15.0
Vermont	6,145	10,503	58.5	5	15	33.3
Virginia	82,390	102,169	80.6	42	87	48.3
Washington	65,552	92,911	70.6	37	90	41.1

TABLE 1 continued

State	NCDB cases	USCS cases	Case coverage, %	CoC facilities	General/surgical facilities[a]	Facility coverage, %
West Virginia	21,230	32,603	65.1	11	57	19.3
Wisconsin	54,856	85,633	64.1	33	129	25.6
Wyoming	2,105	7,085	29.7	0	27	0.0
Missing	5,756	0				
Total	2,890,953	4,292,174	67.4	1455	4750	30.6

USCS United States Cancer Statistics, *NCDB* National Cancer Data Base, *CoC* American College of Surgeons Commission on Cancer

[a] Short-term general and surgical facilities, as reported for 2006 in the Area Resource File

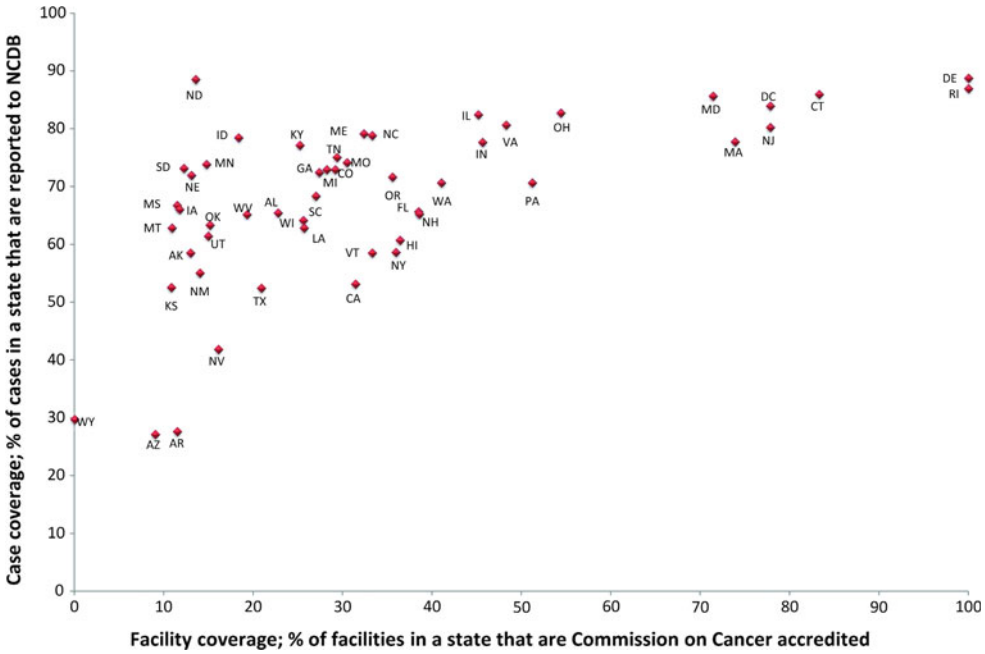


FIG. 1 Scatter plot of states' facility coverage by case coverage, 2004–2006

TABLE 2 USCS and NCDB case counts and coverage by race/ethnicity and age, 2004–2006

Characteristic	USCS	NCDB	Case coverage, %
Race/ethnicity			
White	3,679,070	2,470,154	67.1
Black	431,009	290,693	67.4
American Indian/Alaskan Native	18,559	6,081	32.8
Asian/Pacific Islander	100,019	64,702	64.7
Hispanic[a]	260,619	133,275	51.1
Age			
0–19 years	42,185	30,012	71.1
20–44 years	340,276	263,522	77.4
45–54 years	584,837	435,544	74.5
55–64 years	954,630	673,209	70.5
65–74 years	1,094,511	731,486	66.8
75–84 years	949,079	589,059	62.1
85 + years	326,656	173,877	53.2

USCS United States Cancer Statistics, *NCDB* National Cancer Data Base

[a] Hispanic origin not mutually exclusive from race categories

results not shown), the result of high case coverage for key female sites (e.g., breast, cervix, uterus) and low case coverage for key male sites (e.g., prostate, urinary bladder).

DISCUSSION

To our knowledge, this study is the first to compare case counts in NCDB's hospital-based registry to those in population-based registries. Results suggest that NCDB captured approximately 67 % of newly diagnosed reportable cancer cases. NCDB case coverage varies by state of patient residence, age at diagnosis, and primary site, with smaller differences by race (black, white, Asian/Pacific Islander) and sex. Notable deficiencies in case coverage exist in NCDB with respect to patients of Hispanic ethnicity and Alaskan Natives/American Indians.

As expected, the vast geographic variation in case coverage was highly related to the proportion of hospitals in a given state that were CoC accredited. Two states, Delaware and Rhode Island, had 100 % facility coverage and 88.7 % and 86.9 % case coverage, respectively. Case coverage was not perfectly correlated with facility coverage for a number of reasons; for example, patients may have received treatment out of state, or they may have multiple residences and received treatment closer to a secondary residence. In states where case coverage is high and facility coverage is comparably low (i.e., North Dakota), it is possible that there are a few large facilities providing a full range of clinical services for cancer patients, and that they are CoC accredited. In states where case coverage is low and facility coverage is comparably high, it is possible that one or more large facilities that diagnose and treat a large portion of cancer patients are not CoC accredited.

In NCDB and USCS, Hispanic ethnicity is determined separately from race. State registries contributing data to USCS assign ethnicity using the North American Association of Central Cancer Registries (NAACCR) developed

TABLE 3 USCS and NCDB case counts and coverage by sex and cancer site, 2004–2006

Cancer site	USCS, *n*		NCDB, *n*		Case coverage, %	
	Male	Female	Male	Female	Male	Female
Lung and bronchus	340,477	282,943	223,356	194,705	65.6	68.8
Prostate	615,114		358,060		58.2	
Female breast		591,605		429,151		72.5
Colon and rectum	227,242	219,371	146,886	145,080	64.6	66.1
Urinary bladder	144,887	49,422	88,782	31,450	61.3	63.6
NHL	93,922	82,354	60,573	54,265	64.5	65.9
Melanoma	95,545	72,587	48,058	36,952	50.3	50.9
Kidney and renal pelvis	83,590	52,916	58,242	37,446	69.7	70.8
Uterus, NOS		118,391		91,625		77.4
Leukemias	63,552	48,209	34,886	26,431	54.9	54.8
Benign brain and nervous	39,260	70,395	26,014	46,661	66.3	66.3
Pancreas	52,732	53,592	34,990	35,202	66.4	65.7
Oral cavity and pharynx	67,839	30,515	47,001	21,778	69.3	71.4
Thyroid	22,123	69,794	16,710	53,105	75.5	76.1
Brain and nervous	36,233	29,901	27,818	22,540	76.8	75.4
Ovary		63,660		45,568		71.6
Stomach	38,077	24,460	25,446	16,230	66.8	66.4
Liver and bile duct	39,175	16,091	23,650	9,723	60.4	60.4
Myeloma	28,171	23,655	16,603	13,731	58.9	58.0
Esophagus	35,706	10,351	25,518	7,442	71.5	71.9
Cervix		37,565		29,260		77.9
Larynx	29,551	7,639	20,818	5,786	70.4	75.7
Hodgkin lymphoma	13,679	11,554	9,691	8,340	70.8	72.2
Testis	23,692		16,788		70.9	
Mesothelioma	7,494	2,122	5,173	1,533	69.0	72.2
Kaposi sarcoma	3,518	488	1,849	238	52.6	48.8

USCS United States Cancer Statistics, *NCDB* National Cancer Data Base, *NHL* non-Hodgkin lymphoma, *NOS* not otherwise specified

algorithm commonly known as NHIA (NAACCR Hispanic Identification Algorithm). NHIA uses direct identification (reported in patient charts) as well as indirect identification (a complex algorithm based on last name or family/maiden name). After applying NHIA, any case not identified is coded as non-Hispanic in USCS.[15,16] NHIA has been shown to enhance Hispanic identification in cancer registries by approximately 30 %, on average.[17] NCDB relies only on direct identification; they do not collect protected health information, which includes last name and family/maiden name. Low case coverage among Hispanics may be in part due to USCS's use of NHIA. Another possible reason for low Hispanic case coverage is that Hispanics are less represented in NCDB as a result of geography. According to the U.S. Census Bureau, the top five states by Hispanic population size in 2006 were California, Texas, Florida, New York, and Illinois, with California and Texas vastly outnumbering the other three.[18] Facility coverage in these states ranged from 21 % (Texas) to 45 % (Illinois), and case coverage for California, Texas, and New York were all below 60 %.

American Indians and Alaskan Natives had much lower case coverage compared to other races (32.8 %). Misclassification of American Indian and Alaskan Native race is a major issue faced by all cancer registries; it is further complicated by the reporting of multiple races for a single individual.[19] SEER and NPCR attempt to remedy this misclassification by creating a variable called IHS Link, which associates records in their registries with matching records in the Indian Health Service administrative records database. USCS uses both race and IHS Link to determine American Indian and Alaskan Native race (for Alaska, IHS Link is not used).[15,20] USCS also collect race information from death certificates.[15] Because NCDB does not link to IHS administrative records or obtain information directly from death certificates, it is possible that hospitals reporting to NCDB misclassify American Indian and Alaskan Native race. Geography may again explain some of this difference; Alaska has a very low case coverage (58.5 %), as do four of the five states with the largest American Indian populations (New Mexico 55 %, California 53.1 %, Texas 52.4 %, Arizona 27.1 %).[21]

Low NCDB case coverage is observed among the elderly, which affects overall case coverage because incidence for many common cancers sites peaks in the elderly (e.g., lung, colon, rectum, prostate).[22] There are several potential reasons for this observation. For USCS, no more than 5 % of cancer cases in any state reporting publishable data were ascertained solely on the basis of death certificates. Although these cases are reported to state registries, NCDB does not collect death certificate cases from reporting hospitals, which may help to explain lower case coverage among the elderly.[8] California, Florida, New York, and Texas have the largest elderly populations, yet these states have very low case coverage, ranging from 52.4 % to 65.2 %.[23] Elderly persons encounter more problems related to transportation and health care access compared to other age groups; 21 % of those older than 65 do not drive, and elderly nondrivers make 15 % fewer trips to the doctor.[24] The elderly may be receiving cancer care in smaller, nonurban facilities close to home that are less likely to be CoC accredited, or they may be forgoing care.[1,2,9,10] Although diagnoses at small outpatient facilities would be captured in population-based registries, if the diagnostic facilities are not CoC accredited and the patient does not seek treatment at a CoC-accredited facility as a result of a competing illness or personal preference, it is unlikely these cases will be reported to NCDB. Cancer site may influence case coverage in the elderly; prostate cancer has a very high incidence among the elderly and also low case coverage in NCDB (58.2 %).[22]

Large variations were observed in case coverage by primary site. Because NCDB is a hospital-based cancer registry, one would expect case coverage to be lower for cancer sites that are more frequently diagnosed and treated in outpatient settings (e.g., prostate, melanoma) and higher case coverage for cancers that typically require hospitalization for treatment. Additionally, cancers treated with invasive surgery (e.g., female breast, uterus, esophagus) had higher case coverage than cancers for which surgery is not an integral component of care (e.g., leukemia).

There are some limitations worth noting. A hospital selection bias may exist because NCDB only includes CoC-accredited facilities, which may be systematically different than nonaccredited facilities.[1,2] Because we chose to utilize USCS data as our standard, we could only report NCDB case coverage for variables that USCS makes publicly available. It is not known how inclusion of more recent years of data would affect these estimates, but we hypothesize that the overall case coverage would increase.[25] From diagnosis years 2004 to 2006, the number of CoC facilities reporting data to NCDB has increased from 1345 to 1399. Case coverage similarly increased from 66.7 % in 2004 to 68.4 % in 2006.

The implications of these findings are important to current and future investigators who use NCDB data. The CoC will begin releasing NCDB Participant User Files in 2012 to investigators at CoC-accredited programs who complete an application and project proposal.[26] It is critical that both users of these data and reviewers understand the benefits and limitations of NCDB. It is clear that large variations in NCDB case coverage exist, particularly for certain cancer sites and across geographic regions. As additional facilities seek CoC accreditation, these gaps in case coverage are expected to narrow. These findings could help identify projects for which NCDB data are best suited,

and could also identify potential analyses that may face limitations. Our results also support previous findings that NCDB case coverage is close to 70 % compared to incidence estimates.[4,25] These findings, based on incident cases reported to population-based cancer registries, highlight the strengths of NCDB as a large and important cancer data resource.

ACKNOWLEDGMENT We thank Elizabeth Ward, Stacey Fedewa, and Katherine Virgo from the ACS's Intramural Research Department for their help and guidance with this project. Supported in part by ACS Intramural Research funds.

DISCLOSURE The authors declare no conflict of interest.

REFERENCES

1. American College of Surgeons Cancer Programs. National Cancer Data Base. 2011. http://www.facs.org/cancer/ncdb/index.html. Accessed March 13, 2012.
2. Bilimoria KY, Stewart AK, Winchester DP, Ko CY. The National Cancer Data Base: a powerful initiative to improve cancer care in the United States. *Ann Surg Oncol.* 2008;15:683–90.
3. American College of Surgeons Cancer Programs. About accreditation. 2009. http://www.facs.org/cancer/coc/whatis.html. Accessed August 2, 2012.
4. Jemal A, Murray T, Ward E, Samuels A, Tiwari RC, Ghafoor A, et al. Cancer statistics, 2005. *CA Cancer J Clin.* 2005;55:10–30.
5. Pickle LW, Hao Y, Jemal A, Zou Z, Tiwari RC, Ward E, et al. A new method of estimating United States and state-level cancer incidence counts for the current calendar year. *CA Cancer J Clin.* 2007;57:30–42.
6. Bilimoria KY, Bentrem DJ, Stewart AK, Winchester DP, Ko CY. Comparison of commission on cancer-approved and -nonapproved hospitals in the United States: implications for studies that use the National Cancer Data Base. *J Clin Oncol.* 2009;27:4177–81.
7. United States Cancer Statistics (USCS) Technical Notes: Contributors. 2010. http://www.cdc.gov/cancer/npcr/uscs/2007/technical_notes/contributors/. Accessed March 13, 2010.
8. United States Cancer Statistics (USCS) Technical Notes: USCS publication criteria. 2010. http://www.cdc.gov/cancer/npcr/uscs/2007/technical_notes/criteria.htm. Accessed March 13, 2012.
9. Puts MT, Monette J, Girre V, Wolfson C, Monette M, Batist G, et al. Characteristics of older newly diagnosed cancer patients refusing cancer treatments. *Support Care Cancer.* 2010;18:969–74.
10. Turner NJ, Haward RA, Mulley GP, Selby PJ. Cancer in old age—is it inadequately investigated and treated? *BMJ.* 1999;319(7205):309–12.
11. American College of Surgeons Cancer Programs. Data submission information. 2011. http://www.facs.org/cancer/ncdb/datasubmission.html. Accessed April 19, 2012.
12. Area Resource File (ARF). National County-level Health Resource Information Database. http://arf.hrsa.gov/overview.htm. Accessed April, 2012.
13. United States Cancer Statistics (USCS) Technical Notes: Data sources: incidence data. 2010. http://www.cdc.gov/cancer/npcr/uscs/2007/technical_notes/data_sources/incidence.htm. Accessed March 13, 2012.
14. US Cancer Statistics Working Group. United States cancer statistics, 1999–2007 incidence and mortality Web-based report. Atlanta: US Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute, 2010. http://www.cdc.gov/uscs. Accessed March 13, 2012.
15. United States Cancer Statistics (USCS). Interpreting the data: race and ethnicity in cancer data. 2010. http://www.cdc.gov/cancer/npcr/uscs/2007/technical_notes/interpreting/race.htm. Accessed March 20, 2012.
16. NAACCR Race and Ethnicity Work Group. Springfield, IL: North American Association of Central Cancer Registries, September 2011.
17. Howe HL. Evaluation of NHIA submissions for 1997–2001. Springfield, IL: North American Association of Central Cancer Registries, 2005.
18. Hispanics in the United States: ethnicity and ancestry branch, population division. US Census Bureau, 2006.
19. Espey DK, Wiggins CL, Jim MA, Miller BA, Johnson CJ, Becker TM. Methods for improving cancer surveillance data in American Indian and Alaska Native populations. *Cancer.* 2008;113(5 Suppl):1120–30.
20. Havener L, Hu KU, West DW. Standards for cancer registries, vol. 2: data standards and data dictionary. 10th ed, version 11. Springfield, IL: North American Association of Central Cancer Registries, 2004.
21. Ogunwole SU. The American Indian and Alaska Native population, 2000. US Department of Commerce, Economics and Statistics Administration, US Census Bureau, 2002.
22. Howlader N, Noone AM, Krapcho M, Neyman N, Aminou R, Waldron W, et al. SEER cancer statistics review, 1975–2009 (vintage 2009 populations). 2012. http://seer.cancer.gov/csr/1975_2009_pops09/. Accessed based on November 2011 SEER data submission.
23. US Census Bureau. Statistical abstract of the United States, 2012. 131st ed. 2011. http://www.census.gov/compendia/statab/. Accessed April, 2012.
24. Bailey L. Aging Americans: stranded without options. Washington, DC: Surface Transportation Policy Project, 2004.
25. American College of Surgeons Cancer Programs. Cancer program accreditation. 2011. http://www.facs.org/cancer/coc/approved.html. Accessed April 25, 2012.
26. American College of Surgeons Cancer Programs. Participant user files. 2011. http://www.facs.org/cancer/ncdb/participantuserfiles.html. Accessed March 22, 2012.

Database, Vol. 2012, Article ID bar059, doi:10.1093/database/bar059

Original article

Biocurators and Biocuration: surveying the 21st century challenges

Sarah Burge[1,2], Teresa K. Attwood[3], Alex Bateman[2], Tanya Z. Berardini[4], Michael Cherry[5], Claire O'Donovan[1,2], Ioannis Xenarios[6] and Pascale Gaudet[7,*]

[1]European Bioinformatics Institute, [2]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, [3]Faculty of Life Sciences and School of Computer Science, University of Manchester, Oxford Road, Manchester, M13 9PT, UK, [4]The Arabidopsis Information Resource, Department of Plant Biology, Carnegie Institute for Science, Stanford, CA 94305, [5]Department of Genetics, Stanford University, Stanford, CA 94305-5120, USA, [6]Swiss-Prot Group and [7]CALIPHO Group, Swiss Institute of Bioinformatics, Centre Médical Universitaire, 1, rue Michel Servet CH-1211 Geneva 4, Switzerland

***Corresponding author:** Tel: +41 22 379 5050; Fax: +41 22 379 5858; Email: pascale.gaudet@isb-sib.ch

Submitted 28 September 2011; Revised 25 November 2011; Accepted 28 November 2011

Curated databases are an integral part of the tool set that researchers use on a daily basis for their work. For most users, however, how databases are maintained, and by whom, is rather obscure. The International Society for Biocuration (ISB) represents biocurators, software engineers, developers and researchers with an interest in biocuration. Its goals include fostering communication between biocurators, promoting and describing their work, and highlighting the added value of biocuration to the world. The ISB recently conducted a survey of biocurators to better understand their educational and scientific backgrounds, their motivations for choosing a curatorial job and their career goals. The results are reported here. From the responses received, it is evident that biocuration is performed by highly trained scientists and perceived to be a stimulating career, offering both intellectual challenges and the satisfaction of performing work essential to the modern scientific community. It is also apparent that the ISB has at least a dual role to play to facilitate biocurators' work: (i) to promote biocuration as a career within the greater scientific community; (ii) to aid the development of resources for biomedical research through promotion of nomenclature and data-sharing standards that will allow interconnection of biological databases and better exploit the pivotal contributions that biocurators are making.

Database URL: http://biocurator.org

Introduction

Biocuration involves the analysis, interpretation and integration of biological information into data repositories, primarily to add value by annotating and interconnecting research data and results within a common biological framework. This integration both facilitates the use of data by the wider scientific community and renders them more easily accessible and amenable to computational analysis. A variety of factors, in particular the rapidly increasing pace of data acquisition in the life sciences, have led to the proliferation and wide-spread uptake of biocuration as a full-time career. At the same time, these factors are making the work and careers of biocurators more interesting and challenging every day. Investment in high-throughput technologies, starting with microarray expression analyses in the mid-1990s and continuing with ever improving dissection of the genome, transcriptome, proteome and metabolome, has given rise to a tremendous escalation in the rate of raw biological data production. In turn, this has generated a paradox: on one hand, it has created a pressing need for greater manual annotation and analysis efforts; on the other, it has made it impossible for purely manual efforts to keep up with the scale of data acquisition, creating an urgent need for intelligently designed tools to help automate the conversion of raw data to knowledge and understanding. The challenge for biocurators is clear.

Database, Vol. 2012, Article ID bar059, doi:10.1093/database/bar059

Reflecting the data-driven nature of modern biology, databases have grown considerably both in size and number during the last decade. The exact number of databases is difficult to ascertain. While not exhaustive, the 2011 *Nucleic Acids Research* (*NAR*) online database collection lists 1330 published biodatabases (1), and estimates derived from the ELIXIR database provider survey suggest an approximate annual growth rate of ~12% (2). Globally, the numbers are likely to be significantly higher than those mentioned in the online collection, not least because many are unpublished, or not published in the *NAR* database issue.

Against this background, databases have become a cornerstone of modern biomedical research, and are now being cited in the literature thousands of times per annum. Responsibility for their design, implementation, maintenance, as well as for organizing, annotating, archiving and making their contents publicly available, falls to biocurators and bioinformaticians. As the volumes of data and the number of databases have grown, so too has the biocuration community. In 2009, the International Society for Biocuration (ISB, www.biocurator.org) was formed, to give biocurators a voice and to promote the interests of biocuration. The ISB now counts over 300 members from nearly 150 databases and institutions in 26 different countries. This is a large underestimate: large fractions of the biocuration community are not well-represented in the ISB—in particular, biocurators from commercial databases, as well as researchers, students and post-docs who perform some biocuration work as part of a research project.

While the roles of biocurators in managing and augmenting biomedical data have been increasingly well-documented in the literature (3–6), the nature of their career paths is not well understood, either outside the biocurator community or within it. As part of its mission to advance biocuration as a professional career path, the ISB set out to understand the perceived challenges, concerns and benefits to biocurators of this career choice: specifically, a survey was conducted aiming to gain a qualitative appreciation of biocurators' motivations for entering and remaining in the field, and to comprehend their perceptions of the role of the ISB.

Survey methodology and results

The survey consisted of 37 questions for current biocurators and 13 questions for former biocurators. Questions were a mix of multiple choice, ordinal scale, interval scale and ratio scale. Some questions allowed the respondents to enter a free text reply. There were a total of 257 respondents to the survey. As respondents did not answer every question, the percentages reported correspond to the count of a specific response divided by the total number of responses to a particular question. The survey was publicized through the ISB website, mailing lists and social-networking sites, targeting both current curators and those who had recently left the field. The full survey questions and results are available at (http://biocurator.org/surveys/BiocurationSurveySummary_06292011.pdf). Although the sample size is relatively small, it represents a large fraction of the members of our networks: there are a little over 300 active members in the ISB, over 360 members of the ISB LinkedIn group, and nearly 500 members of the ISB email list (isb@listserv.it.northwestern.edu). Respondents were asked how much time they currently devote to biocuration activities (i.e. up to 10%, up to 50% and up to 100%). The majority of respondents (76%) spent 50–100% of their time on biocuration activities; just over half (53%) were members of the ISB.

Current biocurators

The typical biocurator. Almost 80% of respondents who were currently involved in biocuration were between 31-and 50-years old; 60% were female (Figure 1); and most (71%) were qualified to PhD level. Biocurators come from a range of different scientific backgrounds, most (73%) having previously worked as bench scientists, others (17%) having worked as bioinformaticians, programmers, or in other areas of computational science. Only 11% of respondents described themselves as currently working in industry.

More than half of the respondents (57%) were employed on limited-term contracts, some (25%) of 1–3 years' duration, others (24%) of ≥3 years; 41% were on permanent contracts; and 9% were principal investigators. Notwithstanding the proportion of contract work, 60% of respondents had been in their current role for >4 years, and 82% had been involved in biocuration (in various roles) for ≥7 years.

As shown in Figure 2, the types of data being handled by biocurators were diverse: spanning nucleotide sequences; protein sequences, families, interactions and pathways; small molecules; model organisms; the literature; and

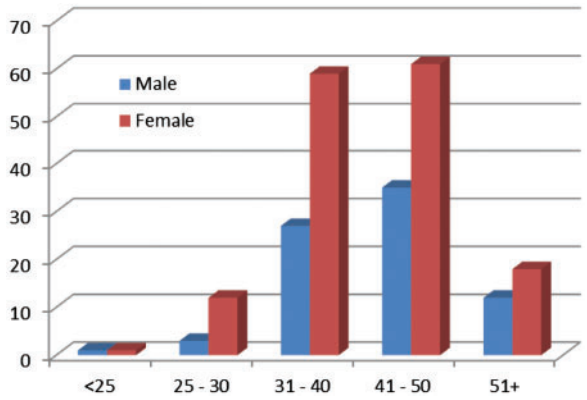


Figure 1. The age and sex distribution of survey respondents.

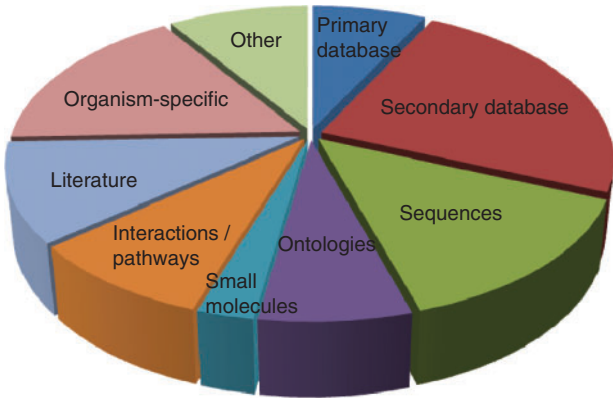


Figure 2. The types of data annotated by biocurators.

Table 1. Main motivations for selecting biocuration as a career

Motivations	%
Wanted to move away from experimental research	48
Intellectual challenge	43
Biocuration is essential for modern science	41
I needed a job	40
Natural transition from previous work	37
The diversity of the work was appealing	33
Previous biocuration experience	11
Knew other biocurators	9
Financial reward	6

biological ontologies. A little over a quarter (26%) of respondents worked with secondary (multi-organism, curated) databases [e.g. UniProtKB (7), IntAct (8), EMAGE (9), 18% with organism-specific resources (e.g. SGD; 10)], dictyBase (11), TAIR (12), etc.]; 16% were involved with sequence analysis; 12% were involved with literature analysis; only 3% worked with chemistry-based resources.

Most biocurators highlighted their appreciation of the team-working aspects of their jobs, with a large fraction collaborating on a regular basis with other scientists, mostly with other curators (78%), and with computer scientists and software engineers (75%); only a minority (17%) of respondents felt isolated in their work. In addition, many respondents were involved in user training/outreach of some sort, whether through posters and talks at meetings (55%), via responses to helpdesk queries (50%), preparation of documentation (45%), or delivery of face-to-face training (34%). For many, these and other activities were associated with some amount of travel, 41% of respondents travelling up to twice a year, and 29% more than 3 times per year. Most (65%) were satisfied with the amount of travelling they did: others (28%) said they would like more travel; 7% wanted less.

Biocuration as a career choice. Respondents were asked what motivated them to become biocurators, by selecting multiple options from a range of pre-set answers. Results are shown in Table 1.

Job satisfaction amongst the surveyed cohort appeared to be high. On a scale of 1–5, where 1 was not at all enjoyable, and 5 was very enjoyable, 77% rated their overall satisfaction at the level of 4 or 5. Moreover, 68% said that they are paid fairly. Nevertheless, only 6% felt that financial reward was a motivating factor in becoming a biocurator. Aspects of the role that particularly contributed to biocurators' enjoyment of their work included the intellectual challenges posed, as well as working extensively with scientific data. Complete results are presented in Table 2.

In the open responses, some respondents highlighted additional benefits they derived from their work, such as enhancement of their analytical thinking, improved ability to critique the literature, and honing of their ability to write concisely.

The survey asked biocurators which aspects of their current work they considered to be important, by ranking the statements on a scale of 1–5 (where 1 was not at all important, and 5 very important); respondents indicated that feeling intellectually challenged, contributing to the direction of the database on which they worked, and keeping abreast of current scientific developments were the most important aspects of their work. They were then asked to consider how strongly they felt that their current roles met these aspirations on a scale of 1–5 (where 1 indicated strong disagreement and 5 as strong agreement). The average answers to those two questions, as well as the differences between the importance of each aspect of the work and how biocuration meets those expectations, are shown in Table 3. Aspects where the differences were largest between expectation and whether biocuration met those expectations were the freedom to choose projects on which biocurators work, as well as recognition from other scientists.

Overall, most respondents derived a sense of accomplishment from their jobs (average, 3.89). Other perceived highlights included learning more and more, and the ability to work remotely, with the consequent lifestyle flexibility that this affords.

These largely positive responses are perhaps reflected in the desire of ~90% of respondents to remain in the field. Nevertheless, 82% expressed concern about future work opportunities, and 60% perceived that the lack of opportunities to move into more senior roles was also a barrier to remain in biocuration. Another concern related to

Database, Vol. 2012, Article ID bar059, doi:10.1093/database/bar059

Table 2. Rewarding aspects of biocuration work

Job aspect	Average rating	Number of 'Enjoyable or very enjoyable' ratings
Intellectual challenges and problem solving	4.39	187
Working with a wide range of scientific data	4.32	183
Working extensively with scientific data	4.21	181
A quantifiable sense of progress	3.87	144
Interaction with end users and data submitters	3.72	130
Scientific work that's not results-driven	3.33	94
Repetitive nature of day-to-day work	2.45	24

Table 3. Important aspects for job satisfaction, and how those aspects are met according to biocurators surveyed

Job aspect	Importance	Job meets expectation	Average difference
Feeling intellectually challenged	4.42	3.88	−0.45
Having an input into the overall direction of your resource	4.27	3.84	−0.36
Keeping abreast of current developments in your scientific area	4.14	3.81	−0.27
Autonomy over work	4.06	3.58	−0.39
Feeling part of a community of scientists	4.06	3.45	−0.48
Recognition from other scientists	3.86	3.30	−0.46
Feeling part of a community of biocurators	3.74	3.24	−0.39
Freedom to choose curation projects	3.73	3.05	−0.54
Freedom to conduct research outside of your core curation responsibilities - both curation-based research and other research	3.36	2.51	−0.70

The numbers presented represent the average score for each aspect, with 1 being the lowest and 5 the highest.

credit: although it was important to feel that their work was recognized by other scientists (average rating, 3.86), there was relatively low confidence that other scientists fully appreciate biocurators' work (average, 3.29).

What makes a good biocurator? Respondents were asked what attributes they thought were important for a biocurator to possess. On a scale of 1–5 (where 1 was not at all important, and 5 very important), respondents indicated that theoretical knowledge (average rating, 4.3), formal scientific training at degree level or above (average, 4.26), good written and verbal communication skills (average, 4.23), and previous experience as an experimental scientist (average, 4.04) were the most important attributes.

On this scale, formal training in data management (average, 2.79) and scripting/programming knowledge (average, 2.58) appear to be less important attributes. However, a significant number of curators did feel that software programming was important, 55% of respondents acknowledging that better training in computer languages would be beneficial, and 43% indicating that they would benefit from better training in bioinformatics. Other aspects that were perceived to be advantageous were improved software (66%), greater automation of routine tasks (61%), and greater adherence to community standards by data submitters (55%).

Career progression

Many biocurators have chosen this career as an alternative to a 'traditional' academic career. However, to attract and retain highly qualified candidates, it is important that opportunities for career progression exist. Although the numbers are far too small to allow us to uncover any trends, 20 of the respondents described themselves as principal investigators, 6 of whom had been in that position for <3 years, which suggests that there are some opportunities for more senior roles within biocuration.

Biocurators who have left the field

The survey also attempted to reach curators who have left the field, to try to gain some understanding of their motivations for doing so. Inevitably, it was challenging to

publicize the survey amongst former biocurators, as they are unlikely to visit the ISB website, or to remain on biocurator-related e-mail lists. Only 10 respondents no longer worked in the field: 9 were aged 41–50 years and held a PhD in a biological science; 6 were women. Two were offered better jobs elsewhere, and although three left because their jobs were no longer sufficiently challenging, five said they had enjoyed their work as a biocurator and five felt that the role had given them general transferable skills or specific skills for their current posts (e.g. critical assessment of publications, and analytical thinking). Three respondents had moved on to work in bioinformatics; four expressed a desire to return to biocuration in the future.

Challenges for the future of biocuration

Respondents were also asked to consider what are the main challenges to biocuration, both by selecting from multiple pre-set answers and with free-text responses. Most (78%) indicated that securing funding to maintain and develop biodatabases was the major threat, and many (71%) also considered that dealing with the increased volumes of data was a significant challenge. Emphasis was also given by many (57%) respondents to the difficulty of impressing on other scientists the importance and hence the need for funding of biocuration. Interestingly, 40% identified with the threat that biocuration might be perceived to be irrelevant if curators cannot keep pace with the current flow of data.

The role of the ISB

The final part of the survey reflected on the role of the ISB in promoting biocuration. Respondents were asked to select those activities they considered most relevant from a range of pre-set answers. Most (88%) felt that the ISB should engage with funding bodies to promote the importance of curation; 80% highlighted the need to engage with journals to encourage the adoption of standard nomenclatures; many (60%) also felt that the ISB should seek out and publicize employment opportunities. Half of the respondents indicated that the ISB should organize and secure funding for regional meetings for curators.

Discussion

Although we recognize that the form of the survey and its results do not lend themselves to rigorous statistical analysis, it has nevertheless yielded some important insights from a fraction of the biocurator community, with respect both to their views on biocuration as a career and to their perceptions of the role of the ISB. Based on the snapshot this survey provided, the career outlook for biocurators seems broadly positive, with high levels of job satisfaction. Respondents generally felt that they benefited from the challenging and problem-solving aspects of their work, yet many highlighted the repetitive nature of the day-to-day job; it is not surprising, therefore, that many respondents highlighted the need for better and more-automated curator-assistant tools, and felt that better training in bioinformatics and software programming would be valuable. Perhaps inevitably, there were concerns about career structure and progression, including the availability of more senior roles and the likelihood of being able to progress into them.

Most of the active biocurators who responded to the survey were >30 years of age. This is consistent with the fact that most respondents held PhDs and had not entered the field directly from their studies, but had held previous posts as bench scientists. This prior experience was clearly considered to be an important attribute for biocurators to possess. Despite the prevalence of contract work, many respondents held permanent posts, and a substantial number had been involved in the field for ≥ 7 years. Biocuration thus appears to lend itself to greater career stability than other scientific fields: the average contract length established for a similar demographic population by the Vitae UK Careers in Research survey was under 3 years, with a majority of those respondents remaining at institutions for <5 years (13).

A secondary aim of this survey was to solicit feedback on curators' perceptions of the role of the ISB in advancing biocuration as a career. Importance was attached to engagement with journals to promote the adoption of standard nomenclature, echoing the view expressed earlier in the survey that adherence to community standards by data submitters would facilitate biocurators' work. The ISB has made substantial commitments to such activities, in collaboration with the BioSharing initiative (14; www.biosharing.org), operating at a global level to build stable linkages between journals and funders, and implementing data-sharing policies and standardization efforts in the biosciences. Members of these two groups have worked in close collaboration with publishers and journals (e.g. Elsevier, Nature Publishing Group, F1000, *Nucleic Acids Research*, *Database*), to develop the BioDBcore standard (15), a proposed uniform system for describing catalogues of databases. Progressively, such efforts will help users to more easily locate and access information dispersed within bio-resources; help shape the data-preservation, data-management and data-sharing policies implemented by journal editors and funders; and encourage software and database developers to embrace and extend community-endorsed standards. In a concrete step towards this goal, BioSharing and the ISB held a workshop at the ISMB meeting in Vienna, in which several journal editors and standards groups stated their commitment to widen participation in, and expedite the implementation of, data-sharing and nomenclature policies

(http://blog.biosharing.org/2011/07/biosharing-at-ismbeccb-2011-vienna.html).

Many respondents also felt that the ISB had a role to play in publicizing employment opportunities and providing opportunities for biocurators to interact with each other. ISB maintains a biocurator job market forum on its website (http://biocurator.org/jobs.shtml), and regularly notifies members of job opportunities via its email lists, through social-networking sites, and the publication of its monthly newsletter (http://biocurator.org/newsletter.shtml). In addition, an international biocuration conference has been held roughly every 18 months since 2005. The ISB has made the support of these conferences part of its mission statement, in order to continue to provide a venue for biocurators and programmers to exchange ideas, discuss their work, improve their methods, and establish collaborations. The Fifth International Biocuration Conference will be held in Georgetown, USA, 2–4 April 2012 (http://pir.georgetown.edu/biocuration2012.html).

Perspective

One concern highlighted by the survey is the possibility that biocuration might, in the future, become irrelevant if biocurators cannot keep up with the onslaught of data. A closely allied fear expressed by a few respondents is the emphasis placed on automatic annotation, and the sense that '[manual] biocuration is meant to be replaced by automated processes'. Given the difficulties of securing funding to support the growing numbers of databases and curators who maintain them, these fears are perhaps understandable. With the pace of data-generation on course to be a million times greater than at present by 2020, there are clearly significant challenges ahead for biocurators. However, without question, the new reality of biological research both demands expert biocurators now in order to make sense of the data deluge, and it assures their role in future, whether at dedicated resources or within research projects; it also argues strongly for continued technological innovation (through deployment of appropriate software, controlled vocabularies, plus data and nomenclature standards) to ensure appropriate use of computers for monotonous high-volume data-processing tasks, releasing biocurators to tackle the current and future intellectual challenges of data management, analysis, interpretation and validation.

It is unfortunate that manual and automatic processes should be considered in opposition, as excluding or superseding each other, or pictured as posing threats to each other. Although many aspects of biocurators' work depend on computation and automation, the development of new tools absolutely requires biologists and bioinformaticians to validate the methods, provide validation tests and ensure their overall usefulness for the community. In addition, several tasks in biocuration can only be performed manually: for instance, the creation of gold standard data sets, and the development of new tools and data models to handle new data types. The research and biocurator communities must work together to ensure that the maximum benefit can be derived from all experimental data being produced. Education of the community on meta-data tagging of data sets, and development of tools to assist with this task, could go a long way to maximizing the utility of data to other researchers. As new areas of biology are explored and new experimental methods are developed, the specific tasks carried out by biocurators may change, but the underlying goal of interpreting, organizing, and making data easily accessible for hypothesis generation and testing will remain essential. The challenges that lie ahead for the biocuration community are not only large, but are also extremely stimulating. We hope that the field will continue to attract innovative and far-sighted scientists to further bridge the gap between data and researchers.

Acknowledgements

We gratefully acknowledge the participation of all the biocurators who responded to the survey.

Funding

Funding for open access charge: International Society for Biocuration.

Conflict of interest. None declared.

References

1. Galperin,M.Y. and Cochrane,G.R. (2011) The 2011 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Res.*, **39 (**Suppl. 1), D1–D6.
2. Southan,C. and Cameron,G. (2009) *Database Provider Survey. Report for ELXIR Work Package 2, July 2009, Version 30 June 2009*, http://www.elixir-europe.org/prep/bcms/elixir/Documents/reports/WP2_Annex-Provider_Survey_Report.pdf.
3. St Pierre,S. and McQuilton,P. (2009) Inside FlyBase: biocuration as a career. *Fly*, **3**, 112–114.
4. Burkhardt,K., Schneider,B. and Ory,J. (2006) A biocurator perspective: annotation at the Research Collaboratory for Structural Bioinformatics Protein Data Bank. *PLoS Comput. Biol.*, **2**, e99.
5. Salimi,N. and Vita,R. (2006) The biocurator: connecting and enhancing scientific data. *PLoS Comp. Biol.*, **2**, e125.
6. Sanderson,K. (2011) Bioinformatics: curation generation. *Nature*, **470**, 295–296.
7. UniProt Consortium. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, **39**(Database Issue), D214–D219.

8. Aranda,B., Achuthan,P., Alam-Faruque,Y. *et al*. (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38**(Database Issue), D525–D531.
9. Richardson,L., Venkataraman,S., Stevenson,P. *et al*. (2010) EMAGE mouse embryo spatial gene expression database: 2010 update. *Nucleic Acids Res.*, **38**(Database Issue), D703–D709.
10. Engel,S.R., Balakrishnan,R., Binkley,G. *et al*. (2010) Saccharomyces Genome Database provides mutant phenotype data. *Nucleic Acids Res.*, **38**(Database Issue), D433–D436.
11. Gaudet,P., Bairoch,A., Field,D. *et al*.; on behalf of the BioDBCore working group (2011a) Towards BioDBcore: a community-defined information specification for biological databases. *Nucleic Acids Res.*, **39**, D7–D10.
12. Swarbreck,D., Wilks,C., Lamesch,P. *et al*. (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, **36**(Database Issue), D1009–D1014.
13. Mellors-Bourne,R. and Metcalfe,J. (2009) Careers in Research Online Survey (CROS) 2009: analysis of aggregated UK results.
14. Field,D., Sansone,S.A., Collis,A. *et al*. (2009) Megascience. 'Omics data sharing. *Science*, **326**, 234–236.
15. Gaudet,P., Fey,P., Basu,S. *et al*. (2011b) dictyBase update 2011: web 2.0 functionality and the initial steps towards a genome portal for the Amoebozoa. *Nucleic Acids Res.*, **39**(Database Issue), D620–D624.

Annals of Surgical Oncology **15**(3):683–690
DOI: 10.1245/s10434-007-9747-3

The National Cancer Data Base: A Powerful Initiative to Improve Cancer Care in the United States

Karl Y. Bilimoria, MD, Andrew K. Stewart, MA, David P. Winchester, MD, and Clifford Y. Ko, MD, MS, MSHS

National Cancer Data Base, Cancer Programs, Commission on Cancer, American College of Surgeons, 636 N. St. Clair Street, 25th Floor, Chicago, IL 60611, USA

The National Cancer Data Base (NCDB) is a nationwide oncology outcomes database that currently collects information on approximately 70% of all new invasive cancer diagnoses in the United States each year and serves as a powerful clinical surveillance and quality improvement mechanism for cancer programs participating in the American College of Surgeons (ACoS) Commission on Cancer (CoC) approvals program.[1–3] Currently, the NCDB receives over one million cancer case reports annually from more than 1,430 hospitals. The NCDB now contains data on more than 21 million cancer patients diagnosed between 1985 and 2005, and is recognized as the largest clinical registry in the world. NCDB data are used to explore trends in cancer care, to examine regional and national benchmarks, and to serve as the basis for quality improvement activities (http://www.facs.org/cancer/ncdb)

The purposes of this review are: (1) to describe the NCDB and the data collected; (2) to discuss how the NCDB can be used to study clinical outcomes and the quality of cancer care in the United States; and (3) to describe the clinical care improvement tools provided by the NCDB and the CoC to participating hospitals.

Received September 19, 2007; accepted November 15, 2007; published online: January 9, 2008.

Address correspondence and reprint requests to: Karl Y. Bilimoria, MD; E-mail: k-bilimoria@northwestern.edu

BACKGROUND: THE COMMISSION ON CANCER AND THE NCDB

The CoC is a multidisciplinary consortium of professional organizations that strive to improve cancer care through setting standards, prevention, research, education, and the monitoring of comprehensive cancer care.[4] The CoC also accredits hospitals as cancer centers, based on their ability to provide a broad range of cancer-related services and specialists.[2] The NCDB was started in 1989 by the ACoS and the CoC in conjunction with the American Cancer Society. The NCDB is housed at the ACoS headquarters in Chicago, Illinois. When the database was established any hospital could voluntarily report data to the NCDB, and during the mid 1990s, more than 2,000 hospitals were contributing cases. In 1996, all CoC-approved hospitals were required to report cancer cases to the NCDB, and in 2001, participation and the associated advantages of reporting to the NCDB were limited to those hospitals who had earned CoC approval.[5]

PARTICIPATING HOSPITALS

The CoC classifies approved hospitals into teaching/research hospitals, comprehensive community cancer centers, and community cancer centers.[2] Teaching/research hospitals must be primarily affiliated with a medical school or National Cancer Institute (NCI)-designated comprehensive cancer center. Community cancer center designations are

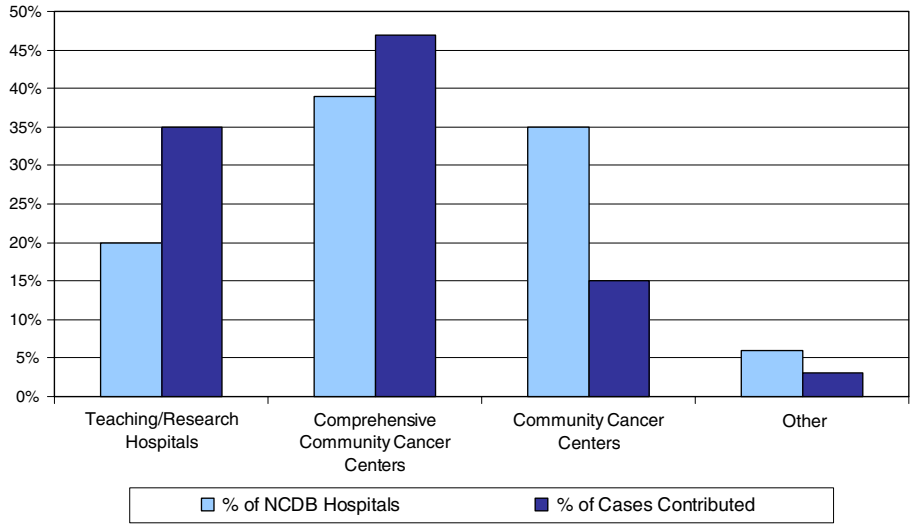


FIG. 1. Percentage distribution of the type of hospitals that report to the NCDB and the percentage of cases contributed by each hospital to the NCDB for 2005. NCDB, National Cancer Data Base.

based on a combination of annual case volume and the ability to provide a wide range of cancer-specific personnel and services. Of the 1,430 hospitals that report to the NCDB, approximately 20% are classified as teaching/research hospitals, 39% are comprehensive community cancer centers, 35% are community cancer centers, and 6% are other cancer centers that include Veterans Health Administration (VA) hospitals (Fig. 1).

Reporting hospitals are located in 49 US states and also Puerto Rico. The cancer programs that report to the NCDB include 19 of 20 adult National Comprehensive Cancer Network (NCCN) hospitals, 33 of 37 NCI-designated cancer centers, and 69 of 121 major inpatient VA hospitals. Some hospitals may elect not to gain CoC approval and thus not contribute data to the NCDB, but the reasons for this are unclear. The 1,430 hospitals that have approved cancer programs and report data to the NCDB account for approximately one third of the inpatient hospitals in the United States. While the NCDB does collect the majority of incident cases in the United States, a hospital selection bias may exist as only CoC-approved hospitals contribute data to the NCDB.

HOW CASES ARE REPORTED

In 1973 federal legislation made invasive cancer a reportable disease, and in 1992 federal legislation further mandated that all cancer cases be reported to state cancer registries.[6] The CoC requires its approved programs to abstract and follow all malignant tumors diagnosed and/or initially treated at the hospital.[7] Annually, CoC-approved hospitals report over one million cancer cases to the NCDB. Approximately 940,000 cases are new invasive cancer diagnoses and account for about 70% of all new cancer diagnoses in the United States, based on annual incidence estimates from the American Cancer Society (Table 1).[8] In addition, the NCDB takes receipt of over 50,000 in situ cancers and 18,000 benign and borderline brain and central nervous system tumors. Patients diagnosed and treated in a physician's office setting, those who undergo consultation to confirm a diagnosis or formulate a treatment plan, and cases reviewed by pathology, but who never enter the hospital for any aspect of their care, are not reported to the NCDB.

Data reporting to the NCDB is highly standardized and similar to other state health departments and federal cancer registry data systems, including the NCI's Surveillance Epidemiology and End Results (SEER) program and the Centers for Disease Control's National Program of Cancer Registries (NPCR).[9,10] Data reported from CoC-approved hospitals are abstracted from patient charts by Certified Tumor Registrars (CTR) who undergo training specific to cancer registry operations.[11] In recent years, the training and certification requirements for CTRs have increased significantly, focusing on many of the nuances of cancer care. If data points are missing, registrars can query the treating physicians to obtain the necessary data to complete the record.

All data submitted to the NCDB undergo a battery of data integrity checks.[12] The rigor of these data

TABLE 1. *Proportion of invasive cancers in the United States reported to the NCDB in 2005 by organ site*

Cancer Site	Number of invasive malignancies reported to the NCDB	Estimated number of annual cancer cases in the United States[1]	Proportion of invasive malignancies reported to the NCDB
Anus	3,476	3,990	87.1%
Bladder[2]	40,647	63,210	64.3%
Bones and joints	2,178	2,570	84.7%
Breast	141,741	212,930	66.6%
Cervix	9,912	10,370	95.6%
Colon	65,576	104,950	62.5%
Esophagus	11,296	14,520	77.8%
Gallbladder and other biliary	6,205	7,480	83.0%
Kidney, renal pelvis, ureter	33,513	38,670	86.7%
Larynx	9,541	9,880	96.6%
Leukemia	20,366	34,810	58.5%
Liver and intrahepatic bile ducts	11,440	17,550	65.2%
Lung and bronchus	141,757	172,570	82.1%
Lymphoma	54,034	63,740	84.8%
Melanoma	28,865	59,580	48.4%
Oral cavity and pharynx	23,824	29,370	81.1%
Other sites	64,503	88,610	72.8%
Ovary	15,225	22,220	68.5%
Pancreas	23,724	32,180	73.7%
Prostate	117,995	232,090	50.8%
Rectum	30,676	40,340	76.0%
Small intestine	4,001	5,420	73.8%
Soft tissue	7,350	9,420	78.0%
Stomach	13,882	21,860	63.5%
Testis	5,729	8,010	71.5%
Thyroid	23,648	25,690	92.1%
Uterus	30,571	40,880	74.8%
TOTAL	941,675	1,372,910	68.6%

[1] Estimated number of invasive malignancies in the United States for 2005 based American Cancer Society estimates.[63] The American Cancer Society has since revised their estimation methodology by basing their calculations on data from state cancer registries rather than on SEER regions. Estimates of cancer incidence have increased with this new methodology.

[2] American Cancer Society estimates for bladder cancer include in situ disease.

Note: The NCDB requires reporting of all in situ cancers for most cancer sites; however, these are not included in the table above. For 2005 diagnoses, more than 50,000 in situ malignancies were reported to the NCDB. The NCDB also receives benign, as well as malignant, brain tumors.

quality-assurance measures was increased significantly for the diagnosis year 2002 and expanded further for 2005 diagnoses. Case records that fail to meet a standardized set of requirements are identified and returned to the hospital. Internal studies comparing the quality of data reporting from large teaching/research hospitals and small community centers revealed no distinct differences in the completeness or accuracy of reported data.[13]

Each year, the NCDB undergoes extensive internal quality monitoring and validity reviews. Aggregate data are examined and verified in advance of the release of annual reports created for each cancer site, and the release of data for analytic purposes.[14] In addition to these extensive automated electronic data-quality checks, site surveyors from the CoC evaluate each hospital's data and data-collection processes once every three years, to ensure that high-quality case information is being reported to the NCDB in an accurate and timely fashion. The NCDB has also undergone multiple data-quality reviews.[15–18]

Once diagnosed and/or treated at a hospital with a CoC-approved cancer program, the patient will be followed by the hospital registrar and the case will be reported to the NCDB regardless of whether the patient receives part of their care at a non-CoC hospital. For example, a patient is diagnosed with a malignancy at hospital A which does not report to the NCDB, undergoes surgery at hospital B which reports to the NCDB, and then returns to hospital A for adjuvant therapy. The CoC requires that the cancer registry at hospital B record the information regarding diagnosis from hospital A and enter follow-up data from hospital A regarding subsequent treatment, recurrence, and survival (active surveillance). If both hospitals were CoC approved, two case records for the same patient would be reported to the NCDB, termed a ‚clinical duplicate.' These duplicate cases are identified using a computer algorithm, and the case

record containing the most complete data is retained in the analytic dataset.

DATA ITEMS COLLECTED

The NCDB collects data on patient factors, tumor characteristics, staging details, surgical and adjuvant treatments, and outcomes, using nationally standardized coding guidelines.[19] All CoC- approved hospitals submitting data to the NCDB are required to use the Facility Oncology Registry Data Standards (FORDS) manual to guide abstracting and coding.[7] Additional guidelines for coding histopathological characteristics and stage of disease are established by the World Health Organization (WHO) in its International Classification of Disease for Oncology (ICD-O) and the American Joint Committee on Cancer (AJCC), respectively.[20,21]

Tumor characteristics include size, the presence and location of nodal and distant metastases, number of nodes positive and examined, grade, margin status, and both clinical and pathological staging information. Clinical stage is a pretreatment assessment that is recorded prior to surgery, and if the patient undergoes resection, a pathological stage is recorded separately. Beginning with cases diagnosed in 2004, the NCDB started collecting Collaborative Stage data elements, in addition to physician-recorded AJCC clinical and pathological tumor–node–metastasis (TNM) elements and overall stage. For the purpose of assessing comorbidity, the International Classification of Disease, 9th Edition (ICD-9-CM), secondary diagnosis codes, excluding cancer diagnoses and post-operative complications, are recorded for each patient.

The NCDB also collects a wide range of data regarding surgical treatment and adjuvant therapy, including radiation, chemotherapy, hormone therapy, and other systemic treatments. Patients are also followed to assess outcomes including readmission within 30 days, recurrence, and survival. CoC-approved hospitals are expected to provide follow-up on at least 90% of all known living patients annually.

Through linkage to tertiary data sources, area-based socioeconomic status and estimates of travel distance/time to the reporting hospital can be obtained based on each patient's home zip code at the time of diagnosis. In addition, hospital type and structural characteristics can be determined through linkage with the CoC's facility information profile system (FIPS) file of self-reported hospital structural features, resources, and services related to oncology.[4]

DATA CONFIDENTIALITY

The ACoS and the NCDB take great care to protect patient and hospital anonymity. The American College of Surgeons has a business associate agreement with each of its CoC-approved hospitals that prohibits the CoC from revealing hospital-specific information. Data reported to the NCDB are retrospective in nature. No direct patient identifiers are collected. Analyses are reported only at the aggregate level to assist hospital cancer programs with quality assurance and improvement. Reported results are in compliance with the privacy requirements of the Health Insurance Portability and Accountability Act (HIPAA).[22]

CLINICAL OUTCOMES STUDIES

The NCDB offers a large, nationally applicable sample to examine oncology practice patterns and outcomes in the United States. Single institutions studies are typically from large centers that focus on particular malignancies. With information from a wide range of hospitals in diverse geographic locations, a better understanding of the variability in care across the country can be analyzed, and, in particular, smaller, non-specialty hospitals can be examined in order to compare differences by hospital type and volume of patients. This is particularly important since the majority of cancer patients in the United States are not treated at high-volume, specialized centers.

Initially, the large number of patients available for analysis through the NCDB resulted in a considerable body of descriptive literature on the clinical and epidemiologic characteristics of common cancers in the United States.[23] Subsequent studies have been more hypothesis driven and have addressed a broad range of cancers and oncological issues. Several studies have used the NCDB to examine surgical and adjuvant therapy treatment trends and utilization.[24–32] The large sample size offered by the NCDB allows studies of rare cancers (i.e., pancreatic neuroendocrine tumors, metaplastic breast cancer, and melanoma in children)[33–38] and subset analyses (i.e., stage I pancreatic cancer patients undergoing resection, 1–2 cm papillary thyroid cancers, T3N0 colon cancer patients)[27,30,39–42] with sufficient statistical power to make reasonable inferences. As AJCC staging information is collected, numerous reports have validated staging systems, and the AJCC has relied heavily on the NCDB to provide data for the

Cancer Staging Manual and to guide staging-system modifications.[21,42–46] Furthermore, studies using NCDB data have addressed important clinical questions in areas where clinical trials may be challenging (i.e., local excision of rectal tumors, and extent of surgery for papillary thyroid cancer).[27,47]

The broad range of hospital types reporting to the NCDB have also facilitated an array of health services research. The NCDB has been used to compare differences in treatment and outcome by hospital type and volume of patients.[25,39,48–50] Data from the NCDB have also been used to examine concordance with established quality measures and standards of clinical care.[26,30,48,51] Multiple studies have also examined cancer care for minorities and disadvantaged populations,[52–54] and a recent article examined how clinical trials could be developed to specifically target minority populations.[55] Statistical methodology pertaining to outcomes research has also been examined for multiple cancer sites.[56] In addition, the NCDB can be used to obtain details beyond those typically collected, by requiring cancer registrars to ascertain additional data points on a specific cohort of patients.[27,50,51,57]

Although there are limitations associated with the use of cancer registry data, appropriate, hypothesis-driven questions can be effectively investigated.[58] Administrative datasets are based on data with financial motivations, such as claims or reimbursement, whereas cancer registries are designed and collect data with the intent of studying cancer epidemiology, treatment trends and utilization, outcomes, and the quality of cancer care. In addition, cautious investigations regarding the efficacy of treatments can also be examined.[27,59] Results from these studies are useful and important in order to generate hypotheses for future prospective studies and clinical trials.

QUALITY IMPROVEMENT

Although the NCDB is a tremendous repository of oncologic data, perhaps the most important aspect is its ability to benchmark hospitals, and feed information back to the public and to participating hospitals. The NCDB offers public benchmark reports for the eleven most commonly diagnosed solid-organ tumors in the United States.[14] These reports currently provide data on more than 5.9 million patients, allowing users to define queries based on patient gender, age, ethnicity, histology, stage, first-course therapy, type of surgical resection, hospital type, and geographic location. The NCDB also offers five-year survival reports stratified by AJCC staging.[60]

In early 2007, the National Quality Forum (NQF) endorsed five quality measures developed in a joint effort between the ACoS, the National Comprehensive Cancer Network (NCCN), and the American Society of Clinical Oncology (ASCO).[61,62] Four of these measures focus on adjuvant therapy for breast and colorectal cancer, and one measure concerns the resection and pathological examination of 12 or more regional lymph nodes for colon cancer. Hospitals with CoC-approved cancer programs reporting to the NCDB are evaluated and benchmarked according to their performance on these measures. Hospitals can examine their performance on each measure in comparison to the other 1,430 CoC hospitals, allowing only the reporting hospital to see where they rank in comparison to the rest of the state, region, or country (Fig. 2). Moreover, this data feedback mechanism allows hospitals to evaluate deficient records, to ensure the accuracy and completeness of their data, and make corrections and additions as needed.

These quality-of-care reports are intended to provide hospitals with the ability to measure and compare standards of cancer care at their facility to that of a large number of hospitals across the country, without any associated accountability. Over the next year, the NCDB will be developing and incorporating additional quality measures for multiple cancers, including common (i.e., lung) and complex (i.e., pancreas) malignancies.

FUTURE DIRECTIONS

The power of the NCDB clearly rests in its ability to receive data, analyze it, and send actionable information back to hospitals for the purposes of quality improvement. These quality improvement activities will play a significant role in the future for the NCDB. In addition to the breast and colorectal initiatives, quality measures must be developed for other cancer sites. These measures should not be used for accountability, but rather only to provide information to hospitals with the aim of improving patient care and outcomes. In the future, federal regulatory agencies, payers, purchasers, and consumers may look to data from cancer registries to assess surgical quality, determine reimbursement levels, and judge providers on the basis of process measures and patient outcomes.

The NCDB will soon have a de-identified dataset available to investigators at CoC-approved hospitals.

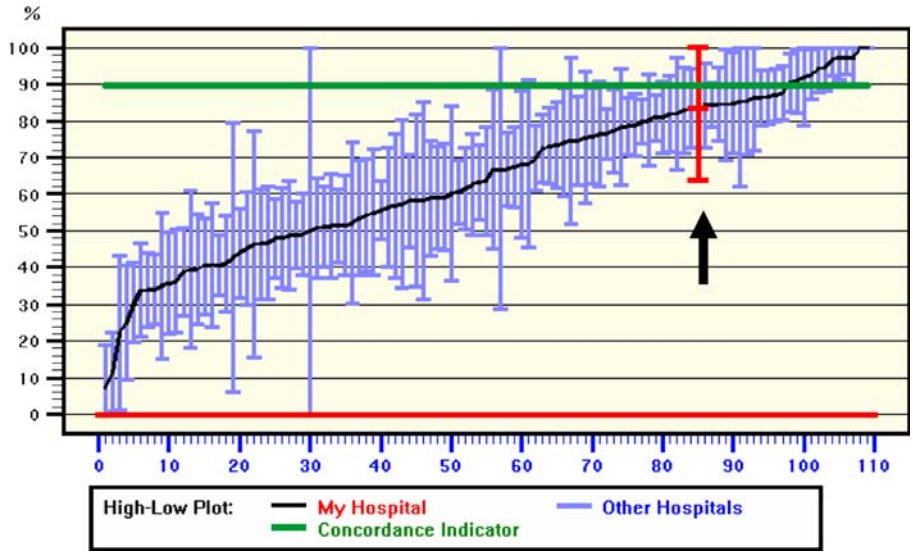


FIG. 2. Sample benchmark report comparing hospitals on the administration/consideration of adjuvant chemotherapy for stage III colon cancer. Note that only the reporting hospital (arrow) can see their individual performance. Results are blinded for all other hospitals. The *x*-axis represents 110 different Commission on Cancer (CoC)-approved hospitals.

This file will provide the opportunity to examine a large dataset containing a number of data items that are not available through other cancer registries including surgical margin status, systemic therapy (chemotherapy and hormonal therapy), detailed radiation treatment data, and clinical and pathological AJCC staging.

CONCLUSIONS

The NCDB is a large, powerful database providing multiple opportunities for clinical studies and quality-improvement endeavors. Importantly, the NCDB can be used to benchmark hospitals on performance measures and serve as an impetus for quality improvement initiatives at the hospital level. Cancer registration activities are expensive for hospitals, but the benefits are clearly demonstrated by the NCDB's extensive efforts to feed important information back to participating institutions. The NCDB is the only cancer registry that provides feedback data to hospitals in such a manner. It is the responsibility of surgeons, and their colleagues in other oncologic specialties, to work collectively, and with their hospital's cancer registrars, to continuously review and improve the accuracy, completeness and quality of the hospital's cancer registry data, and thus the quality of the data reported to the NCDB. The NCDB is a valuable tool to improve the quality of cancer care in the United States.

ACKNOWLEDGEMENTS

The authors would like to recognize the staff of the National Cancer Data Base: James M. Banasiak, E. Greer Gay, Kathy Mallin, Bryan Palis, Florin Petrescu, Jerri Linn Phillips, and Kimberly R. Porter.

REFERENCES

1. National Cancer Data Base. http://www.facs.org/cancer/ncdb/index.html. Cited 15 May 2007.
2. Commission on Cancer: Approvals Categories. http://www.facs.org/cancer/coc/categories.html. Cited 16 July 2007.
3. Winchester DP, Stewart AK, Bura C, et al. The National Cancer Data Base: A clinical surveillance and quality improvement tool. *J Surg Oncol* 2004; 85:1–3.
4. Commission on Cancer.http://www.facs.org/cancer. Cited 15 July 2007.
5. Commission on Cancer. Cancer Program Approval. http://www.facs.org/cancer/coc/whatis.html. Cited 1 July 2007.
6. Cancer Registries Amendment Act, Public Law 102–515. Congressional Record 1992; 138, 106 STAT 3372–7.
7. Facility Oncology Registry Data Standards. Chicago: Commission on Cancer; 2004.
8. Jemal A, Murray T, Ward E, et al. Cancer statistics, 2005. *CA Cancer J Clin* 2005; 55:10–30.
9. National Cancer Institute: SEER Surveillance Epidemiology and End Results. http://seer.cancer.gov/. Cited 26 January 2007.
10. National Program of Cancer Registries, Cancer Surveillance System. http://www.cdc.gov/cancer/npcr/training/css.htm. Cited 11 August 2007.
11. National Cancer Registrars Association: Certified Tumor Registrar certification. http://ncra-usa.org/certification/index.htm. Cited 12 July 2007.
12. Centers of Disease Control and Prevention: National Program of Cancer Registries, EDITS. http://www.cdc.gov/cancer/npcr/tools/edits/. Cited 1 July 2007.
13. A. K. Stewart, E. G. Gay, L. Patel-Parekh, et al. Provider feedback improves reporting on quality measures: National

profile reports for adjuvant chemotherapy for stage III colon cancer. Presented at the American Society of Clincal Oncology Annual Meeting, Chicago, 2007.
14. NCDB Public Benchmark Reports. http://www.facs.org/ncdbbenchmarks8.cfm. Cited 1 July 2007.
15. Sylvester M, Blankenship C, Carter A, et al. Quality control: The American College of Surgeons Commission on Cancer Standards, National Cancer Data Base, and Cancer Liason Program. *J Reg Mgmt* 2001; 28:68–74.
16. Eberle C, Phillips J, Tary P, et al. Quality management in the National Cancer Data Base: A re-abstracting study of the Midwest region. *J Reg Mgmt* 1997; 24:93–7.
17. Fremgen A, Jessup J, Menck HR. Hospital use of NCDB data to assess quality of cancer patient care. *J Reg Mgmt* 1995; 22:69–71.
18. Mulnar K, Phillips J, Fritz A, et al. Quality of oncology data: Findings from the Commission on Cancer PCE study. *J Reg Mgmt* 2001; 28:24–34.
19. Standards for Cancer Registries. Data Standards and Data Dictionary, 8th Ed., Version 10.1, Vol. II. Springfield, IL: North American Association of Central Cancer Registries; 2003.
20. International Classification of Disease for Oncology, 3rd ed. Geneva: World Health Organization; 2000.
21. AJCC Cancer Staging Manual. 6th ed. Chicago, IL: Springer; 2002.
22. Health Insurance Portability and Accountability Act (HIPAA), 1996. Standards for Privacy of Individually Identifiable Health Information; Final Rule (45 CFR Parts 160 and 164).
23. National Cancer Data Base: Bibliography. http://www.facs.org/cancer/ncdb/biblclin.html. Cited 12 July 2007.
24. Bilimoria KY, Tomlinson JS, Merkow RP, et al. Clinicopathologic features and treatment trends of pancreatic neuroendocrine tumors: analysis of 9,821 patients. *J Gastrointest Surg* 2007; 11:1460–9.
25. Bilimoria KY, Bentrem DJ, Linn JG. Utilization of total thyroidectomy for papillary thyroid cancer in the United States. *Surgery* 2007; 142(6):906–13.
26. Jessup JM, Stewart A, Greene FL, et al. Adjuvant chemotherapy for stage III colon cancer: implications of race/ethnicity, age, and differentiation. *J Am Med Assoc* 2005; 294:2703–11.
27. You YN, Baxter NN, Stewart A, et al. Is the increasing rate of local excision for stage I rectal cancer in the United States justified? A nationwide cohort study from the National Cancer Database. *Ann Surg* 2007; 245:726–33.
28. David KA, Milowsky MI, Ritchey J, et al. Low incidence of perioperative chemotherapy for stage III bladder cancer 1998 to 2003: A report from the National Cancer Data Base. *J Urol* 2007; 178:451–4.
29. Bilimoria KY, Bentrem DJ, Ko CY, et al. Multimodality therapy for pancreatic cancer in the U.S.: Utilization, outcomes, and the effect of hospital volume. *Cancer* 2007; 110:1227–34.
30. Kennedy T, Stewart AK, Bilimoria KY, et al. Treatment trends and factors associated with survival in T1aN0 and T1bN0 breast cancer patients. *Ann Surg Oncol* 2007; 14:2918–27.
31. Hoffman HT, Porter K, Karnell LH, et al. Laryngeal cancer in the United States: Changes in demographics, patterns of care, and survival. *Laryngoscope* 2006; 116:S1–13.
32. Singletary SE, Patel-Parekh L, Bland KI. Treatment trends in early-stage invasive lobular carcinoma: A report from the National Cancer Data Base. *Ann Surg* 2005; 242:281–9.
33. Bilimoria KY, Talamonti M, Tomlinson JS, et al. Prognostic score predicting survival after resection of pancreatic neuroendocrine tumors: Analysis of 3,851 patients. *Ann Surg* 2007; in press.
34. Pezzi CM, Patel-Parekh L, Cole K, et al. Characteristics and treatment of metaplastic breast cancer: Analysis of 892 cases from the National Cancer Data Base. *Ann Surg Oncol* 2007; 14:166–73.
35. Lange JR, Palis BE, Chang DC, et al. Melanoma in children and teenagers: an analysis of patients from the National Cancer Data Base. *J Clin Oncol* 2007; 25:1363–8.
36. Zhen W, Karnell LH, Hoffman HT, et al. The National Cancer Data Base report on squamous cell carcinoma of the base of tongue. *Head Neck* 2004; 26:660–74.
37. Smith RB, Apostolakis LW, Karnell LH, et al. National Cancer Data Base report on osteosarcoma of the head and neck. *Cancer* 2003; 98:1670–80.
38. Lee JH, Barich F, Karnell LH, et al. National Cancer Data Base report on malignant paragangliomas of the head and neck. *Cancer* 2002; 94:730–7.
39. Bilimoria KY, Bentrem DJ, Ko CY, et al. National failure to operate on early stage pancreatic cancer. *Ann Surg* 2007; 246:173–80.
40. Bilimoria KY, Bentrem DJ, Ko CY, et al. Validation of the 6th edition AJCC Pancreatic Cancer Staging System: report from the National Cancer Database. *Cancer* 2007; 110:738–44.
41. Swanson RS, Compton CC, Stewart AK, et al. The prognosis of T3N0 colon cancer is dependent on the number of lymph nodes examined. *Ann Surg Oncol* 2003; 10:65–71.
42. Fong Y, Wagman L, Gonen M, et al. Evidence-based gallbladder cancer staging: Changing cancer staging by analysis of data from the National Cancer Database. *Ann Surg* 2006; 243:767–71; discussion 771–4.
43. Bilimoria K, Bentrem DJ, Merkow R, et al. Application of the pancreatic adenocarcinoma staging system to pancreatic neuroendocrine tumors. *J Am Coll Surg* 2007; in press.
44. Bilimoria KY, Bentrem DJ, Ko CY, et al. Validation of the 6(th) edition AJCC pancreatic cancer staging system: report from the National Cancer Database. *Cancer* 2007; 110:738–44.
45. Greene FL. TNM staging for malignancies of the digestive tract: 2003 changes and beyond. *Semin Surg Oncol* 2003; 21:23–9.
46. Greene FL, Stewart AK, Norton HJ. New tumor-node-metastasis staging strategy for node-positive (stage III) rectal cancer: An analysis. *J Clin Oncol* 2004; 22:1778–84.
47. Bilimoria KY, Bentrem DJ, Ko CY, et al. Extent of surgery affects survival for papillary thyroid cancer. *Ann Surg* 2007; 246:375–384.
48. Bilimoria K, Stewart A, Palis B, et al. Adequacy and importance of nodal examination for colon cancer in the elderly. J Am Coll Surg 2007.
49. Bilimoria K, Talamonti MS, Ko C, et al. Effect of hospital type and volume on lymph node evaluation for gastric and pancreatic cancer. Western Surgical Association Annual Meeting, Colorado Springs, Colorado, 2007.
50. Reid-Lombardo KM, Gay G, Patel-Parekh L, et al. Treatment of gastric adenocarcinoma may differ among hospital types in the United States, a report from the National Cancer Data Base. *J Gastrointest Surg* 2007; 11:410–9; discussion 419–20.
51. Malin JL, Kahn KL, Adams J, et al. Validity of cancer registry data for measuring the quality of breast cancer care. *J Natl Cancer Inst* 2002; 94:835–44.
52. Halpern MT, Bian J, Ward EM, et al. Insurance status and stage of cancer at diagnosis among women with breast cancer. *Cancer* 2007; 110:403–11.
53. Shin SH, Palis BE, Perry RR. Cutaneous Melanoma in Asian Americans. Society of Surgical Oncology 2007 Annual Meeting, Washington, DC, 2007.
54. McGinnis LS, Menck HR, Eyre HJ, et al. National Cancer Data Base survey of breast cancer management for patients from low income zip codes. *Cancer* 2000; 88:933–45.
55. Newman LA, Lee CT, Parekh LP, et al. Use of the National Cancer Data Base to develop clinical trials accrual targets that are appropriate for minority ethnicity patients: a report from the American College of Surgeons Oncology Group (ACOSOG) Special Population Committee. *Cancer* 2006; 106:188–95.

56. Bilimoria KY, Stewart A, Tomlinson JS, et al. Impact of adjuvant radiation on survival: A note of caution when using cancer registry data to evaluate adjuvant treatments. *Ann Surg Oncol* 2007; in press.
57. Miller DC, Spencer BA, Ritchey J, et al. Treatment choice and quality of care for men with localized prostate cancer. *Med Care* 2007; 45:401–9.
58. Warren JL, Harlan LC. Can cancer registry data be used to study cancer treatment?. *Med Care* 2003; 41(9):1003–5.
59. Bilimoria KY, Bentrem DJ, Ko CY, et al. Extent of surgery affects survival for papillary thyroid cancer. *Ann Surg* 2007; 246:375–81; discussion 381–4.
60. NCDB Public Survival Reports. http://www.facs.org/ncdbbr/survival8.cfm. Cited 1 July 2007.
61. Commission on Cancer (CoC). http://www.facs.org/cancer/qualitymeasures.html. Cited 26 April 2007.
62. Quality of Cancer Care Performance Measures. National Quality Forum website. http://216.122.138.39/projects/ongoing/cancer/index.asp. Cited 17 January 2007.
63. American Cancer Society, Cancer Facts and Figures. http://www.cancer.org/docroot/STT/content/STT_1x_Cancer_Facts__Figures_2007.asp. Cited 21 January 2007.