Business Case Model for use of PDQ in Data Capture QA

A White Paper by K. Bradley Paxton, Ph.D.
15 Jun 2011

Executive Summary

In doing forms data capture, whether with just human data entry keying from paper forms or with Optical Character Recognition (OCR), Optical Mark Recognition (OMR) and Key From Image (KFI), it has been customary to employ manual methods for data quality assurance. These methods involve a process we refer to as Double Key & Verify (DK&V), wherein one keyer is asked to key a particular data field, and then another keyer is asked to key the same field (preferably without collusion). If the results from both keyers agree with the sampled field, then the sampled data field is deemed to be correct. If they do not, then a third party is usually employed to divine the correct answer. This classic DK&V process is slow and costly; so, in practice, the amount of data sampled for Quality Assurance (QA) purposes is often smaller than desired for useful statistically valid results.

At ADI, we have developed a new automated approach to data capture QA that we call Production Data Quality (PDQ). In brief, the PDQ system employs independent data capture engines to sample the production data capture images and results, and to quickly and cost-effectively determine the truth of the sampled fields. By cost-effective, we mean that the amount of human keying needed for QA may be reduced by a factor of 40 or more. In addition, when the truth is known, data quality accuracy can be measured precisely, root-cause analysis is enabled, and data capture system improvements made more rapidly.

A version of this PDQ system was developed and used in the U.S. Census Bureau's 2010 Decennial Census and was a great success, providing near-real-time feedback to Census management of data capture issues and assurance that the required data quality metrics were being met.

This short white paper describes a useful business case model to use to estimate the keyer QA savings you can achieve using PDQ. It is not our intent here to explain the internal workings of PDQ in detail, but a process flow chart is shown in Appendix 1. The model's nominal values herein are built around the CMS-1500 Health Insurance Claim Form, examples of which are shown in Appendices 2 and 3, however, the model can be used for any form type you wish by simply changing the model inputs appropriately.

How to Use the Business Case Model

Using the model discussed in this white paper is easy to do. Basically, you just put in values that describe your form and data entry process in the input section; and the savings to you if you were to employ PDQ to do your data entry QA are immediately calculated. If you want to try a different estimate of a particular value, you can easily change it and see "what if."

Below we will briefly describe the inputs, and then show you an example. Table 1 below lists the ten inputs and some nominal values.

Assumptions (Inputs)	Values
Form Type	CMS-1500
OCR Accept Rate	0.8
Keystrokes/Hour	6000
Burdened Keying Cost ($/Hour)	15
Form Volume (Megaforms)	1
Characters/Form	1000
OCR QA Sampling Rate	0.01
KFI QA Sampling Rate	0.05
DK&V Factor	2
PDQ Efficiency over DK&V	40

Table 1 – The Ten Inputs to the PDQ Business Case Model
(With some typical values)

Form Type

Here you just record the name of the form type from which you are capturing data, just to label the calculation so you can remember what you did. Here, in our example, we have assumed the form type is the CMS-1500 Health Insurance Claim Form, formerly referred to as "HCFA."

OCR Accept Rate

The OCR Accept Rate is that fraction of your data capture work being read automatically by your recognition system. Here, we assume 0.8 (or 80%). The remainder of the work, in this case 0.2 (or 20%) is the Reject Rate, and this production work is sent to your human keyers because the OCR is unsure about the answer. If you are doing all human data entry keying and not using automation at all, then simply enter 0.0 for the OCR Accept Rate (equivalent to rejecting everything to keyers).

Keystrokes/Hour

This is just the average number of keystrokes you estimate your keyers punch per hour under normal daily (not peak) working conditions. Here, we have assumed 6,000 keystrokes per hour.

Burdened Keying Cost ($/Hour)

This one is a little tricky, because it needs to be the total cost of employing a keyer in your enterprise. Ideally, it would include not only the hourly wage, but also the cost of equipment, space, heat and light, etc. You may need your CPA. Here, we have assumed an approximate U.S. minimum wage with a 2X burden rate, or about $15/hour.

Form Volume (Megaforms)

Here we use a new term, "Megaforms," which as you can probably guess means a volume of one million forms. We introduced this term some time back, as the volume at which data capture automation should seriously be considered. Here, we assume only one "Megaform," which is also easy to mentally scale up.

Characters/Form

This input is the average number of characters placed on a form by the respondent, whether machine print or handprint. We have estimated 1000 characters based on a CMS-1500 form.

OCR QA Sampling Rate

This would be the rate at which you wish to statistically sample the (accepted) output from your OCR system for QA purposes. Here, we have assumed a sampling rate of 0.01 (or 1%), which was what was actually done in Census 2010.

KFI QA Sampling Rate

This is the rate at which you wish to sample the results of your human keying for QA purposes, whether Key From Paper (KFP) or Key From Image (KFI). Here, we assume 0.05 (or 5%), which is what was actually done in Census 2010.

DK&V Factor

This factor just accounts for the extra keying required when Double Key & Verify is employed. We have only assumed a factor of two here, but often it is really about 2.2 in practice due to the "verify" part.

PDQ Efficiency over DK&V

This number can really be from about 30 for very discriminating studies like the Census up to 100 or more, depending on the overall accuracy of your system that you are trying to measure. You can think of it as the reduction factor for keying effort in doing the QA by employing PDQ instead of DK&V. Here, we assume a nominal value of 40 (you may envision your army of QA keyers being divided by 40 if you use PDQ). In order for you to really believe this one, you may have to try it and see how it works with your actual forms and your workflow, but this is where the money is.

The "Live" PDQ Business Case Model

If we start by assuming that all the above assumptions are OK for now, and run the model, then you get the following:

Assumptions (Inputs)	Values
Form Type	CMS-1500
OCR Accept Rate	0.8
Keystrokes/Hour	6000
Burdened Keying Cost ($/Hour)	15
Form Volume (Megaforms)	1
Characters/Form	1000
OCR QA Sampling Rate	0.01
KFI QA Sampling Rate	0.05
DK&V Factor	2
PDQ Efficiency over DK&V	40

Results (Outputs)	Values
Total Character Volume	1.00E+09
OCR Character Volume	8.00E+08
KFI Character Volume	2.00E+08
QA Character Volume	3.60E+07
QA Keying Time in Hours	6.0E+03
QA Cost using DK&V	$90,000
QA Cost using PDQ	$2,250

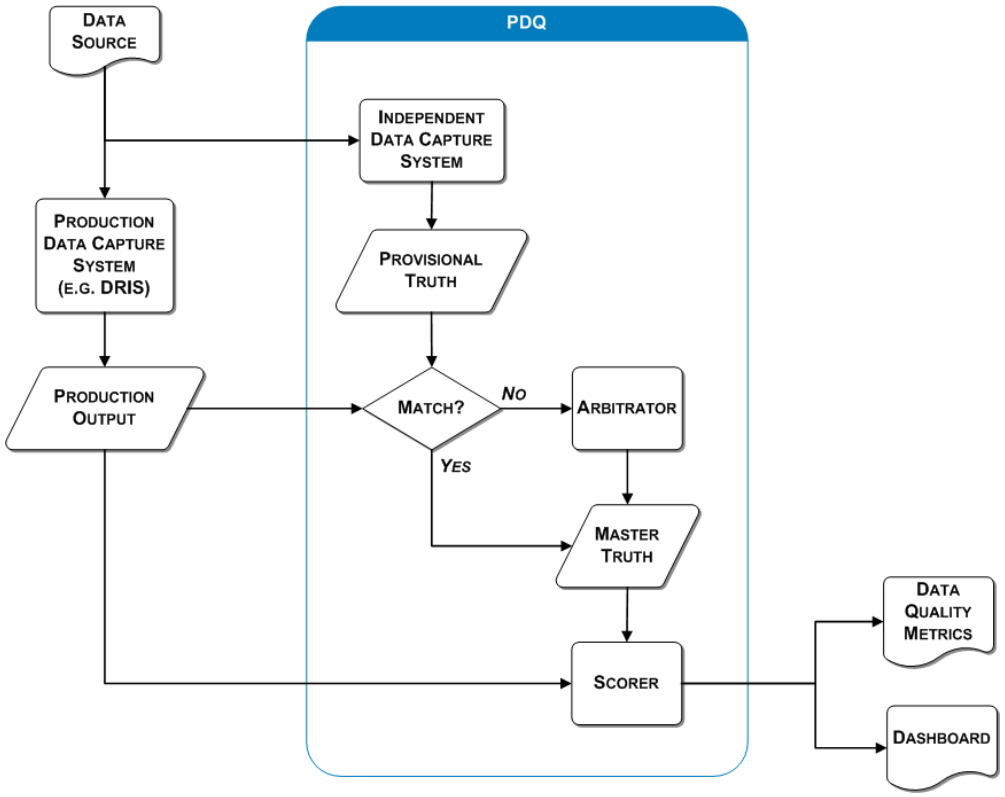
QA Savings using PDQ	**$87,750**

This white paper is just focused on the obvious QA cost savings, but you can also make significant continuous quality improvements using PDQ, which will be the topic of another paper.

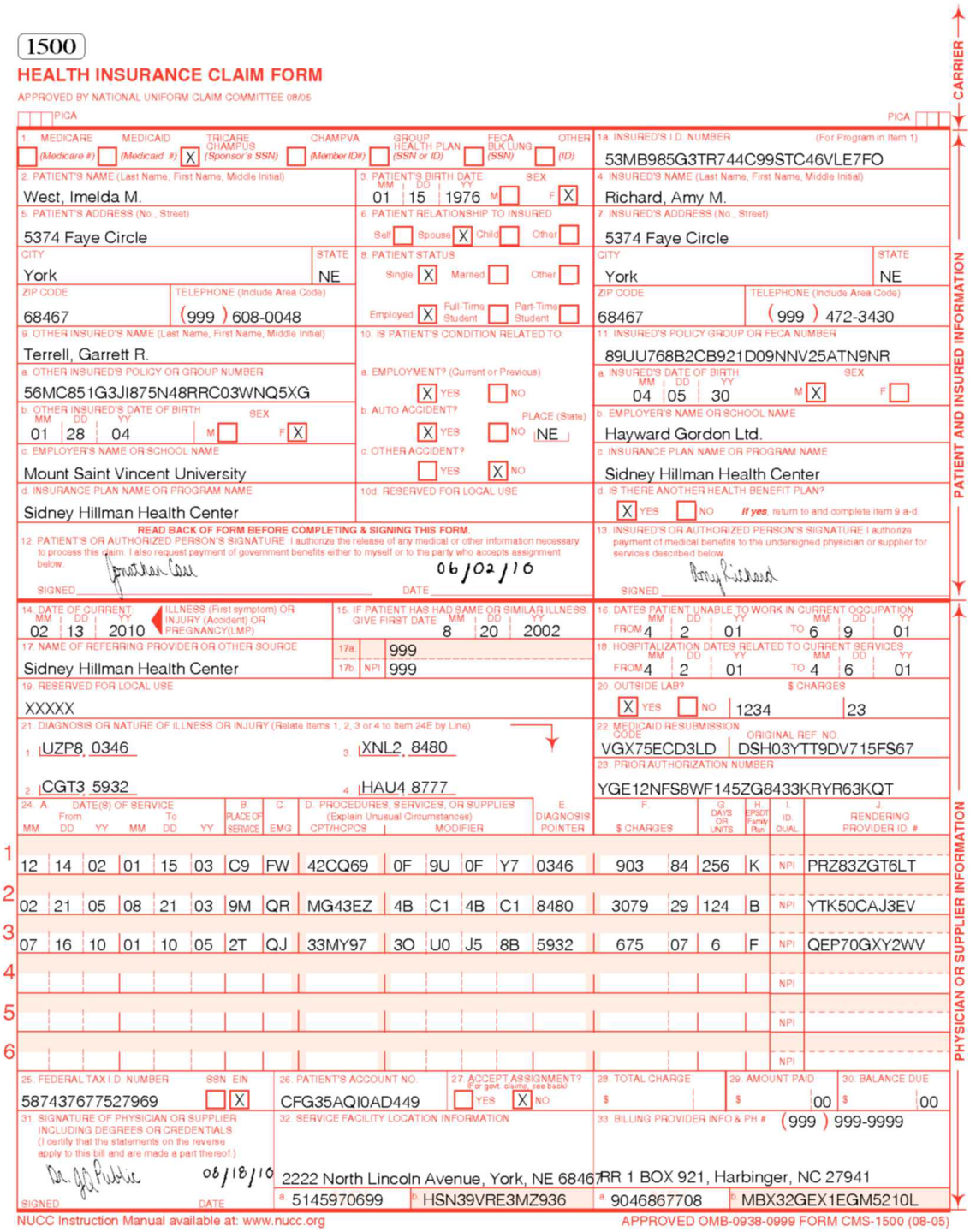
For further information, or to get an Excel workbook of this model, contact:
Brad Paxton
ADI, LLC
200 Canal View Boulevard, Suite 100
Rochester, NY 14623
(585) 239-6057
brad.paxton@adillc.net
http://www.adillc.net

Appendix 1 – PDQ Process Flow Diagram

We have intentionally not tried to explain the inner workings of PDQ in this note that is focused just on QA cost savings. However, here is a high-level flow chart of PDQ in case you're interested.



Appendix 2 – Example of a Machine-Printed CMS-1500 Form (Synthetic Data)



1500

HEALTH INSURANCE CLAIM FORM

APPROVED BY NATIONAL UNIFORM CLAIM COMMITTEE 08/05

PICA

1. MEDICARE (Medicare #) | MEDICAID (Medicaid #) | TRICARE CHAMPUS (Sponsor's SSN) | CHAMPVA (Member ID#) | GROUP HEALTH PLAN (SSN or ID) | FECA BLK LUNG (SSN) | OTHER (ID) — MEDICAID [X]

1a. INSURED'S I.D. NUMBER (For Program in Item 1): 53MB985G3TR744C99STC46VLE7FO

2. PATIENT'S NAME (Last Name, First Name, Middle Initial): West, Imelda M.

3. PATIENT'S BIRTH DATE: 01 15 1976 SEX: F [X]

4. INSURED'S NAME (Last Name, First Name, Middle Initial): Richard, Amy M.

5. PATIENT'S ADDRESS (No., Street): 5374 Faye Circle

6. PATIENT RELATIONSHIP TO INSURED: Self | Spouse [X] | Child | Other

7. INSURED'S ADDRESS (No., Street): 5374 Faye Circle

CITY: York — STATE: NE

8. PATIENT STATUS: Single [X] | Married | Other

CITY: York — STATE: NE

ZIP CODE: 68467 — TELEPHONE (Include Area Code): (999) 608-0048

Employed [X] | Full-Time Student | Part-Time Student

ZIP CODE: 68467 — TELEPHONE (Include Area Code): (999) 472-3430

9. OTHER INSURED'S NAME (Last Name, First Name, Middle Initial): Terrell, Garrett R.

10. IS PATIENT'S CONDITION RELATED TO:

11. INSURED'S POLICY GROUP OR FECA NUMBER: 89UU768B2CB921D09NNV25ATN9NR

a. OTHER INSURED'S POLICY OR GROUP NUMBER: 56MC851G3JI875N48RRC03WNQ5XG

a. EMPLOYMENT? (Current or Previous): [X] YES | NO

a. INSURED'S DATE OF BIRTH: 04 05 30 SEX: M [X] | F

b. OTHER INSURED'S DATE OF BIRTH: 01 28 04 SEX: M | F [X]

b. AUTO ACCIDENT? [X] YES | NO — PLACE (State): NE

b. EMPLOYER'S NAME OR SCHOOL NAME: Hayward Gordon Ltd.

c. EMPLOYER'S NAME OR SCHOOL NAME: Mount Saint Vincent University

c. OTHER ACCIDENT? YES | [X] NO

c. INSURANCE PLAN NAME OR PROGRAM NAME: Sidney Hillman Health Center

d. INSURANCE PLAN NAME OR PROGRAM NAME: Sidney Hillman Health Center

10d. RESERVED FOR LOCAL USE

d. IS THERE ANOTHER HEALTH BENEFIT PLAN? [X] YES | NO — *If yes*, return to and complete item 9 a-d.

READ BACK OF FORM BEFORE COMPLETING & SIGNING THIS FORM.

12. PATIENT'S OR AUTHORIZED PERSON'S SIGNATURE I authorize the release of any medical or other information necessary to process this claim. I also request payment of government benefits either to myself or to the party who accepts assignment below.

SIGNED: Jonathan Lee — DATE: 06/02/10

13. INSURED'S OR AUTHORIZED PERSON'S SIGNATURE I authorize payment of medical benefits to the undersigned physician or supplier for services described below.

SIGNED: Amy Richard

14. DATE OF CURRENT: ILLNESS (First symptom) OR INJURY (Accident) OR PREGNANCY(LMP): 02 13 2010

15. IF PATIENT HAS HAD SAME OR SIMILAR ILLNESS, GIVE FIRST DATE: 8 20 2002

16. DATES PATIENT UNABLE TO WORK IN CURRENT OCCUPATION: FROM 4 2 01 TO 6 9 01

17. NAME OF REFERRING PROVIDER OR OTHER SOURCE: Sidney Hillman Health Center

17a. 999

17b. NPI 999

18. HOSPITALIZATION DATES RELATED TO CURRENT SERVICES: FROM 4 2 01 TO 4 6 01

19. RESERVED FOR LOCAL USE: XXXXX

20. OUTSIDE LAB? [X] YES | NO — $ CHARGES: 1234 23

21. DIAGNOSIS OR NATURE OF ILLNESS OR INJURY (Relate Items 1, 2, 3 or 4 to Item 24E by Line)

1. UZP8 0346
2. CGT3 5932
3. XNL2 8480
4. HAU4 8777

22. MEDICAID RESUBMISSION CODE: VGX75ECD3LD — ORIGINAL REF. NO.: DSH03YTT9DV715FS67

23. PRIOR AUTHORIZATION NUMBER: YGE12NFS8WF145ZG8433KRYR63KQT

	24. A. DATE(S) OF SERVICE From MM DD YY	To MM DD YY	B. PLACE OF SERVICE	C. EMG	D. PROCEDURES, SERVICES, OR SUPPLIES (Explain Unusual Circumstances) CPT/HCPCS	MODIFIER	E. DIAGNOSIS POINTER	F. $ CHARGES	G. DAYS OR UNITS	H. EPSDT Family Plan	I. ID. QUAL.	J. RENDERING PROVIDER ID. #
1	12 14 02	01 15 03	C9	FW	42CQ69	0F 9U 0F Y7	0346	903	84	256	K	NPI PRZ83ZGT6LT
2	02 21 05	08 21 03	9M	QR	MG43EZ	4B C1 4B C1	8480	3079	29	124	B	NPI YTK50CAJ3EV
3	07 16 10	01 10 05	2T	QJ	33MY97	3O U0 J5 8B	5932	675	07	6	F	NPI QEP70GXY2WV
4												NPI
5												NPI
6												NPI

25. FEDERAL TAX I.D. NUMBER: 587437677527969 — SSN | EIN [X]

26. PATIENT'S ACCOUNT NO.: CFG35AQI0AD449

27. ACCEPT ASSIGNMENT? (For govt. claims, see back) YES | [X] NO

28. TOTAL CHARGE: $

29. AMOUNT PAID: $ 00

30. BALANCE DUE: $ 00

31. SIGNATURE OF PHYSICIAN OR SUPPLIER INCLUDING DEGREES OR CREDENTIALS (I certify that the statements on the reverse apply to this bill and are made a part thereof.)

SIGNED: Dr. JJ Fiebie — DATE: 08/18/10

32. SERVICE FACILITY LOCATION INFORMATION: 2222 North Lincoln Avenue, York, NE 68467

a. 5145970699 — b. HSN39VRE3MZ936

33. BILLING PROVIDER INFO & PH #: (999) 999-9999

RR 1 BOX 921, Harbinger, NC 27941

a. 9046867708 — b. MBX32GEX1EGM5210L

NUCC Instruction Manual available at: www.nucc.org

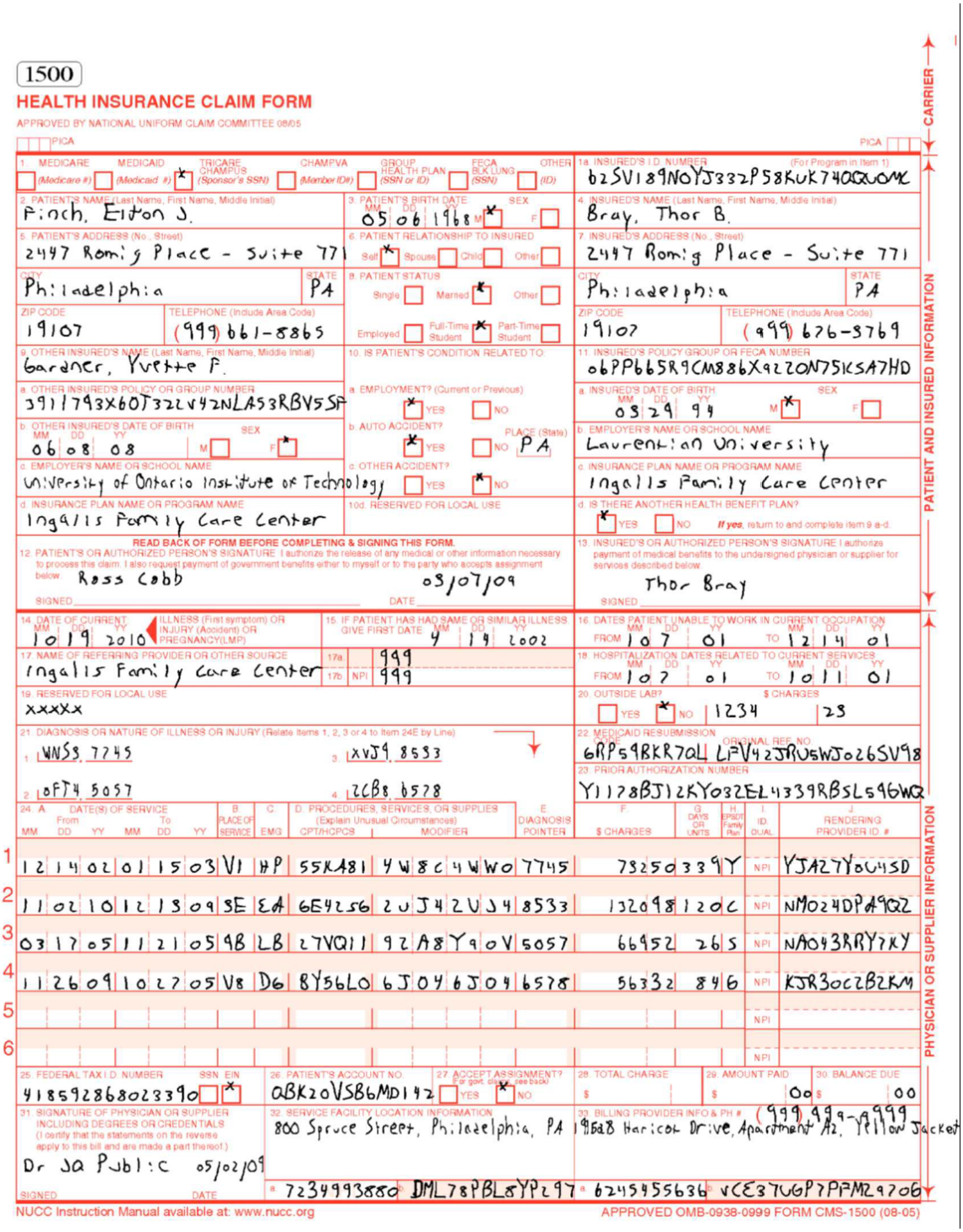
APPROVED OMB-0938-0999 FORM CMS-1500 (08-05)

CARRIER

PATIENT AND INSURED INFORMATION

PHYSICIAN OR SUPPLIER INFORMATION

Appendix 3 – Example of a Hand-Printed CMS-1500 Form (Synthetic Data)



1500

HEALTH INSURANCE CLAIM FORM

APPROVED BY NATIONAL UNIFORM CLAIM COMMITTEE 08/05

PICA

1. MEDICARE (Medicare #) MEDICAID (Medicaid #) TRICARE CHAMPUS (Sponsor's SSN) CHAMPVA (Member ID#) GROUP HEALTH PLAN (SSN or ID) FECA BLK LUNG (SSN) OTHER (ID)

1a. INSURED'S I.D. NUMBER (For Program in Item 1): b2SV189NOYJ332P58KUK74QQUOMC

2. PATIENT'S NAME (Last Name, First Name, Middle Initial): Finch, Elton J.

3. PATIENT'S BIRTH DATE: 05 06 1968 SEX: M

4. INSURED'S NAME (Last Name, First Name, Middle Initial): Bray, Thor B.

5. PATIENT'S ADDRESS (No., Street): 2447 Romig Place - Suite 771

6. PATIENT RELATIONSHIP TO INSURED: Self

7. INSURED'S ADDRESS (No., Street): 2447 Romig Place - Suite 771

CITY: Philadelphia STATE: PA

8. PATIENT STATUS: Married

CITY: Philadelphia STATE: PA

ZIP CODE: 19107 TELEPHONE (Include Area Code): (999) 661-8865

Employed: Full-Time Student

ZIP CODE: 19107 TELEPHONE (Include Area Code): (999) 676-9769

9. OTHER INSURED'S NAME (Last Name, First Name, Middle Initial): Gardner, Yvette F.

10. IS PATIENT'S CONDITION RELATED TO:

11. INSURED'S POLICY GROUP OR FECA NUMBER: 06PPB65R9CM886X92Z2ON75ICSA7HD

a. OTHER INSURED'S POLICY OR GROUP NUMBER: 391179 3X60T32ZV42NLA53RBV5SF

a. EMPLOYMENT? (Current or Previous): YES

a. INSURED'S DATE OF BIRTH: 05 29 94 SEX: M

b. OTHER INSURED'S DATE OF BIRTH: 06 08 08 SEX: M

b. AUTO ACCIDENT? YES PLACE (State): PA

b. EMPLOYER'S NAME OR SCHOOL NAME: Laurentian University

c. EMPLOYER'S NAME OR SCHOOL NAME: University of Ontario Institute of Technology

c. OTHER ACCIDENT? NO

c. INSURANCE PLAN NAME OR PROGRAM NAME: Ingalls Family Care Center

d. INSURANCE PLAN NAME OR PROGRAM NAME: Ingalls Family Care Center

10d. RESERVED FOR LOCAL USE

d. IS THERE ANOTHER HEALTH BENEFIT PLAN? YES If yes, return to and complete item 9 a-d.

READ BACK OF FORM BEFORE COMPLETING & SIGNING THIS FORM.

12. PATIENT'S OR AUTHORIZED PERSON'S SIGNATURE I authorize the release of any medical or other information necessary to process this claim. I also request payment of government benefits either to myself or to the party who accepts assignment below. SIGNED: Ross Cobb DATE: 05/07/09

13. INSURED'S OR AUTHORIZED PERSON'S SIGNATURE I authorize payment of medical benefits to the undersigned physician or supplier for services described below. SIGNED: Thor Bray

14. DATE OF CURRENT: ILLNESS (First symptom) OR INJURY (Accident) OR PREGNANCY(LMP): 10 19 2010

15. IF PATIENT HAS HAD SAME OR SIMILAR ILLNESS, GIVE FIRST DATE: 04 19 2002

16. DATES PATIENT UNABLE TO WORK IN CURRENT OCCUPATION: FROM 10 07 01 TO 12 14 01

17. NAME OF REFERRING PROVIDER OR OTHER SOURCE: Ingalls Family Care Center

17a. 999

17b. NPI 999

18. HOSPITALIZATION DATES RELATED TO CURRENT SERVICES: FROM 10 07 01 TO 10 11 01

19. RESERVED FOR LOCAL USE: xxxxx

20. OUTSIDE LAB? NO $ CHARGES: 1234 23

21. DIAGNOSIS OR NATURE OF ILLNESS OR INJURY (Relate Items 1, 2, 3 or 4 to Item 24E by Line)

1. WN55 7745
2. 0FT4 5057
3. XVJ9 8553
4. ZCB8 6578

22. MEDICAID RESUBMISSION CODE: ORIGINAL REF. NO.: 6RP59BKR7QL LFV42JRU6WJ026SV98

23. PRIOR AUTHORIZATION NUMBER: Y1178BJ12KY032EL4339RBSL596WQ

24. A. DATE(S) OF SERVICE From MM DD YY	To MM DD YY	B. PLACE OF SERVICE	C. EMG	D. PROCEDURES, SERVICES, OR SUPPLIES (Explain Unusual Circumstances) CPT/HCPCS	MODIFIER	E. DIAGNOSIS POINTER	F. $ CHARGES	G. DAYS OR UNITS	H. EPSDT Family Plan	I. ID. QUAL.	J. RENDERING PROVIDER ID. #
1	12 14 02 01 15 03	V1	HP	55X481	4W8C4WW07745		73250	33	9Y	NPI	YJA27Y0U4SD
2	11 02 10 12 13 09	3E	EA	6E4256	2UJ42UJ48533		13209	8	12	0C NPI	NM024DPA9QZ
3	03 17 05 11 21 05	9B	LB	27VQ11	92A8Y90V5057		66952	26	5	NPI	NA043RRY7XY
4	11 26 09 10 27 05	V8	D6	8Y56LO	6J04 6J04 6578		56332	84	6	NPI	KJR3OC2B2KM
5										NPI	
6										NPI	

25. FEDERAL TAX I.D. NUMBER: 418592868023390 SSN EIN: EIN

26. PATIENT'S ACCOUNT NO.: QBK20VSB6MD142

27. ACCEPT ASSIGNMENT? (For govt. claims, see back) NO

28. TOTAL CHARGE: $

29. AMOUNT PAID: $ 00

30. BALANCE DUE: $ 00

31. SIGNATURE OF PHYSICIAN OR SUPPLIER INCLUDING DEGREES OR CREDENTIALS (I certify that the statements on the reverse apply to this bill and are made a part thereof.) Dr Ja Public 05/02/09 SIGNED DATE

32. SERVICE FACILITY LOCATION INFORMATION: 800 Spruce Street, Philadelphia, PA

a. 7234993880 b. DML78PBL8YP297

33. BILLING PROVIDER INFO & PH #: (999 999-9999 19688 Haricot Drive, Apartment A2, Yellow Jacket

a. 6245455636 b. VCE37UGP7PPMZ9706

NUCC Instruction Manual available at: www.nucc.org

APPROVED OMB-0938-0999 FORM CMS-1500 (08-05)

CARRIER

PATIENT AND INSURED INFORMATION

PHYSICIAN OR SUPPLIER INFORMATION

Yes, this one is a bit challenging, but that was done on purpose for a client.