Estimation of the Size of the Media Necessary to Construct a Medical Image Database

Mitsuru Ikeda[1], Takeo Ishigaki,[2] and Kazunobu Yamauchi[1]

[1]Department of Medical Information and Medical Records,

Nagoya University Hospital.

[2]Department of Radiology, Nagoya University, School of Medicine.

(Address)

Department of Medical Information and Medical Records

Nagoya University Hospital

65, Tsurumai-cho, Showa-ku, Nagoya 466, JAPAN.

Abstract

To estimate the size of the media necessary to construct a medical image database, a statistical analysis was made of the radiological data in our hospital information system database. The distribution of the total radiological image data storage required during one working day can be regarded as a normal distribution, and the distribution per patient was different from a normal one. Therefore, the mean amount required for the data storage of radiological images during one working day is very useful in estimating the data storage required for radiological images during a given number of working days.

Key words: Picture archiving and communication system (PACS).
Model analysis.
System assessment.
Hospital information system (HIS).
Radiology and radiologists.

1. Introduction

The Picture Archiving and Communication System (PACS) has been reported to have substantial merits [1]. However, there are only a few clinically acceptable PACS in operation. One of the major problems in constructing PACS is that the image data storage capacity required for the database is very large. Therefore, estimating this capacity is very important in designing a practical clinical image database system for PACS; this estimation is also useful for estimating the cost of managing digital diagnostic images [2-4].

However, statistical analysis of the data storage required for medical information including images in the database has been superficial so far. Therefore, a statistical analysis was made of the radiological data in our hospital information system (HIS) database. We also estimated the size of the media necessary to store most of the radiological images from a 922-bed academic hospital in the PACS database.

2. Materials

For this study, a statistical analysis was made of the radiological examination data in the Nagoya University Hospital Information System called CHART [5, 6]; this analysis covered the 12 months from October 1, 1993, to September 30, 1994. During that period, 24,868 patients underwent radiological examinations. For our present analysis, we collected data from radiological orders using the data retrieval system developed in our HIS; this data retrieval system has been installed in the UNIX workstations connected through LAN to the mainframe computer of our HIS [7].

3. Calculation of Image Data Storage Required

In this present study, we calculated the radiological image data storage required if all radiological images were stored on optical disks and were handled by the image database. However, nuclear medicine examinations, ultrasound, and magnetic resonance images were excluded from this calculation, because the data on these examinations in the HIS database did not have the information necessary for this calculation. The following algorithm was used for calculating the radiological (digital) image data storage required:

(1) The general formula for calculating the bytes needed to store image data on the optical disks is

$$(\text{Bytes of one image storage}) = H + X' \times Y \times B \times CR,$$

in which H is a Header size required for image identification, X' is derived from a pixel array size for a horizontal axis X by

$$X' = \begin{cases} \left([X/U] + 1\right) \times 128 & (\text{if } (X \bmod U) \neq 0) \\ X & (\text{if } (X \bmod U) = 0) \end{cases},$$

Y is a pixel array size for a vertical axis, B is bytes of storage required for each pixel, U is a natural number determined by the minimum data storage unit of the optical disk (for our calculations, U is 128), and CR is a compression ratio.

(2) To estimate the amount required for storage of radiological examinations using conventional X-ray film (plain X-ray examinations, angiography, urography, myelography, and so on), it is assumed that the amounts of such digitized information correspond to those in the Fuji computed radiography system (FCR; Fuji Medical Systems, Tokyo)

(3) The digital array sizes ($X \times Y$) for imaging plates in Fuji computed radiography are 1) 1760×2140 (for 14"×17"), 2) 1760×1760 (for 14"×14"), 3) 1670×2010 (for 10"×12"), and 4) 2000×2510 (for 8"×10"); and X''s are 1) 1792 (for 14"×17"), 2) 1792 (for 14"×14"), 3) 1792 (for 10"×12"), and 4) 2048

(for 8"×10"). Each pixel requires 2 storage bytes ($B = 2$) and the compression ratio is 0.4 ($CR = 1/2.5$). In addition, 12228 storage bytes are required for header information ($H = 12288$).

(4) The digital array size ($X \times Y$) of one image from the CT examination is 512×512, and X' is 512. Each pixel requires 2 storage bytes ($B = 2$). Images from the CT examination are stored without compression. In addition, 512 storage bytes are required for header information ($H = 512$).

(5) The digital array size ($X \times Y$) of one image of the digital subtraction angiography (DSA) is 1024×1024, and X' is 1024. Each pixel requires 2 storage bytes ($B = 2$). DSA images are stored without compression. In addition, 2908 bytes of storage are required for header information ($H = 2908$).

Figure 1 shows the distribution of total radiological (digital) image data storage required per patient over 12 months; the mean of these was 30.32 ± 97.44 Mbytes. This is not a normal distribution. The distribution of total radiological (digital) image data storage required during one working day is shown in Figure 2; the mean of these was 2.97 ± 0.78 Gbytes. This is similar to a normal distribution. Therefore, we made the least square fit of these frequencies (as a function of total radiological image data storage required during one working day T) to the Gauss function, $a \exp\left[-\frac{(T-m)^2}{2\sigma^2}\right]$. Here, a, m, and σ^2 are parameters for which the following estimated values were obtained: $a = 25.9 \pm 1.2$, $m = 2.91 \pm 0.02$, $\sigma^2 = 0.141 \pm 0.015$ (the correlation coefficient for these parameters R^2 is 0.928).

4. Estimation of Image Data Storage Required during a Given Number of Working Days

To estimate the total radiological image data storage required during a given number of working days, we calculated the mean of 20 kinds of totals of the randomly sampled amounts required for the data storage of radiological images during one working day for each number of working days (1-245); that is, for a given number of working days (D days), $\frac{\sum_{j=1}^{20}\left(\sum_{i=1}^{D} T_{random(245)}\right)}{20}$ was calculated, where, $T_i\ (i = 1,\cdots 245)$ is the total radiological image data storage required during one working day, and *random*(245) is the function generating random numbers in the range 1 through 245.

These simulated amounts required for the data storage of radiological images ($T(D)$) were plotted as a function of the number of working days (Figure 3). We made the least square fit to the linear function, $T(D) = T_m \times D$, where T_m is a parameter, and the estimated value for T_m was 2.9547 ± 0.0002 with a correlation coefficient R^2 of 1.000. The simulated amounts approximated the actual ones.

5. Estimation of Image Data Storage Required per Given Number of Patients

In addition, to estimate these amounts per given number of patients, we calculated the mean of 20 kinds of totals of the randomly sampled amounts required for the data storage of radiological images during the 12 months per patient for each number of patients (100-24800, every 100), that is, for a given number of patients (P patients), $\frac{\sum_{j=1}^{20}\left(\sum_{i=1}^{P} S_{random(24868)}\right)}{20}$ was calculated, where $S_i\ (i = 1,\cdots 24868)$ is the total radiological image data storage required during the

12 months per patient, and *random*(24868) is the function generating random numbers in the range 1 through 24868.

These simulated amounts required for the data storage of radiological images during the 12 months $S(D)$ were plotted as a function of the number of patients (Figure 4). We made the least square fit to the linear function, $S(D) = b \times D$, where b is a parameter, and the estimated value for b was 0.027850 ± 0.000003 with a correlation coefficient R^2 of 1.000. However, the degree of the approximation of simulated amounts per given number of patients to the actual ones was not as good as that of the simulated ones during a given number of working days.

We also calculated the following three means of 20 kinds of totals of randomly sampled amounts of the data storage of radiological images per patient for each number of patients: the first one was calculated using data from October 1, 1993, to December 31, 1993; the second using data from October 1, 1993, to March 31, 1994; and the third using data from October 1, 1993, to June 31, 1994. These simulated amounts required for data storage of radiological images were also plotted as a function of the number of patients (Figure 4); these relationships were linear with correlation coefficients of 1.000. The slopes estimated by the least square fit to the linear relationships increased linearly as the analyzing period increased (its correlation coefficient R^2 was 0.945).

6. Calculation of Data Storage Required for Medication and Laboratory Test Results

Moreover, a statistical analysis was made of medication data and laboratory test results in the database of our HIS; the period of this analysis covered the 3 months from January 1, 1994, to March 31, 1994. We calculated the amount required for storage of all these data. The following algorithm was used for calculating the amount required for storage of medication data and laboratory test results:

(1) To estimate medication data storage, we assume that the data storage required for one prescribed drug item is 36 bytes; this information includes the drug identification code, the prescription date, the dose, the usage identification code, and so on.

(2) In estimating the data storage of laboratory test results, we assume that one item of a laboratory test result requires 18 bytes; this information includes the laboratory test identification code and test result.

The amount required for medication data storage during these 3 months was 10.104 Mbytes, and that for laboratory test results was 22.977 Mbytes. The amount required for data storage of radiological images during the same 3 months was 188.161 Gbytes. (During these 3 months, there were 61 working days.) Thus, the amount required for storage of medication data and laboratory test results was only 0.012-0.005% of that for radiological images.

To obtain a correlation among the amounts required for data storage of medical information per patient, the Pearson correlation coefficient (r) was calculated for patients with medication data, laboratory test results, and radiological data stored in our HIS database for the 3 months from January 1, 1994, to March 31, 1994. There was a weak correlation between the amounts (per patient) required for storage of medication data and laboratory test results ($r = 0.404$, $p<0.001$). However, the amounts required for storage of radiological image data per patient showed no correlation with those required for storage of medication data and laboratory test results ($r = 0.131$ and $r = 0.332$, respectively).

7. Discussion

Although some studies have been made to estimate the data storage required of radiological digital images [1, 3], little is known about their statistical characteristics. In these studies, the estimation of data storage required of the radiological data was derived from the estimation of data storage requirements for a single working day [3], but such a derivation may present problems. The estimated amount of radiological image data during one working day may be the mean of these. However, the mean value is not very useful, if the distribution of the amounts is not a normal one. Therefore, we have investigated the statistical characteristics of the amount required for data storage of radiological images. Our investigations reveal that the distribution of total radiological image data storage required during one working day can be regarded as a normal one, and that the distribution of total radiological image data storage required per patient is different from a normal distribution.

However, the above results (that is, the distribution of total radiological image data storage required during one working day can be regarded as a normal distribution) may be incidental, since these amounts may change slightly according to the calculation methods. With that in mind, we also calculated the distributions of numbers of total radiological images during one working day. It follows from these calculations that those distributions are similar to normal distributions for almost modalities. For example, Figure 5 shows the distribution of the numbers of all $14'' \times 14''$ size X-ray films used for radiological examinations during one working day; the mean of these was 240 ± 68. This distribution is also similar to a normal one, and we made the least square fit of these frequencies (as a function of the numbers of all X-ray films during one working day N) to the Gauss function, $a \exp\left[-\frac{(N-m)^2}{2\sigma^2}\right]$, where, a, m, and σ^2 are parameters for which the following estimated values were obtained: $a = 28.2 \pm 1.8$, $m = 235.8 \pm 5.3$, $\sigma^2 = 4901 \pm 740$. The correlation coefficient for these parameters R^2 was 0.892. The linear combination of the random variables due to normal distributions and independent of each other is also due to a normal distribution. Therefore, regardless of the calculation algorithm, the distribution of total radiological image data storage required during one working day can be regarded as a normal distribution. These results are probably due to the fact that the limited capacity to perform radiological examinations in the hospital may determine the number of radiological examinations during one working day.

The distribution of the amounts required for data storage of radiological digital images during one working day (T_i) can be regarded as a normal one. Therefore, the distributions of the totals of these amounts ($\sum_{i=1}^{D} Ti$) for a given number of working days (D days) can be also regarded as normal, because the sum of the random variables due to normal distributions and independent of each other is also due to a normal distribution. Furthermore, since the mean of $\sum_{i=1}^{D} Ti$ is given by (the mean of T_i) × D, the mean amount required for data storage of radiological images during one working day is very useful in estimating the amount required for such storage during a given number of working days ($T(D)$). In fact, the estimated total radiological image data storage required during a given number of working days was almost equal to the actual amount. In addition, the mean amount during one working day was also in good agreement with the estimated value of parameter T_m from the least square fit to $T(D) = T_m \times D$.

From the above, the estimate of the size of the media required for storage of radiological image data during a given number of working days (D days) from an academic 922-bed hospital is given by $2.955 \times D$ Gbytes. Furthermore, the estimate of the media size required for D days data is generally given by $T_m \times D$

(where, T_m is the mean amount required for storage of such data during one working day).

On the other hand, the distribution of the total radiological image data storage required per patient during 12 months (S_i) was different from the normal one; in this distribution, there was a peak around the mean, and the frequencies after the peak gradually decreased. Therefore, some errors in the estimation of the amount required for the data storage of radiological images for a given number of patients arise from the postulation that probability distributions of these amounts in one patient are normal distributions. Hence, the mean amount required for data storage of radiological images per patient is not very useful in estimating the amount required for storage of radiological image data for a given number of patients.

However, an estimation of data storage of radiological images per patient is necessary to construct databases specifically designed for research applications in radiation therapy or interventional radiology. From our results in such cases, the size of the media necessary to construct the database can be also given approximately by $S_m \times P$ (where, S_m is the mean of the amounts required for storage of radiological data during the period required for the database user per patient, and P is the number of patients which should be stored in the database). However, this estimation has some errors which are especially great when the estimation period is different from the one during which the mean amount of data storage of images per patient is calculated, and when the linear complement must be used for the estimation.

From our results, the amount required for storage of medical information data other than image data was negligible, compared with that for image data. (Further, the amounts required for storage of radiological image data per patient did not correlate with those for storage of medication data and laboratory test results.) Therefore, the estimation of the size of the media necessary to construct the global image database combined with clinical information is approximately given by $T_m \times D$.

8. Conclusions

The distribution of the amount required for the data storage of radiological digital images during one working day can be regarded as a normal one. On the other hand, the distribution of the total radiological image data storage required per patient was different from the normal distribution. Hence, the mean amount required for data storage of radiological images during one working day is very useful in estimating the amount required during a given number of working days. The amount required for storage of medical information other than image data was negligible. Therefore, the size of the media necessary to construct the global image database combined with clinical information is given approximately by $T_m \times D$.

References

1. U. P. Schmiedl and A. H. Rowberg, Literature review: picture archiving and communication systems, *Journal of Digital Imaging* **3**, 178-194 (1990).

2. K. G. Vosburgh, Storage and retrieval of radiographic images, *Radiology* **123**, 619-624 (1977).

3. S. J. Dwyer III, A. W. Templeton, N. L. Martin, K. R. Lee, E. Levine, S. Batnitzky, S. J. Rosenthal, D. F. Preston, H. I. Price, S. Faszold, W. H. Anderson, and L. T. Cook, The cost of managing digital diagnostic images, *Radiology* **144**, 313-318 (1982).

4. E. M. S. J. V. Gennip and A. R. Bakker, Challenge and opportunities for technology assessment in medical informatics. case study: PACS, *MED. INFORM.* **18**, 209-218 (1993).

5. K. Yamauchi, Y. Suzuki, M. Ikeda, and T. Miura, Total hospital information system using an optical disk filing system for medical record management, *Proceedings of the Seventh World Congress on Medical Informatics* 255-259 (1992).

6. K. Yamauchi, M. Ikeda, Y. Suzuki, M. Asai, K. Toyama, and E. Hayashi, Evaluation of the order entry system by end users - a step to the new hospital information system, *Nagoya J. Med. Sci.* **57**, 19-24 (1994).

7. M. Ikeda, E. Hayashi, and K. Yamauchi, Model analysis of time duration in a medication order entry system with attention to do-medication orders, *Comput. Biol. Med.* **24**, 473-483 (1994).

Figure Legends

1. **Figure 1.**

The distribution of total radiological image data per patient during 12 months.

2. **Figure 2.**

The distribution of total radiological image data during one working day. The curved line indicates the least square fit of these frequencies to the Gauss function.

3. **Figure 3.**

Graph depicts simulated and actual total radiological image data as a function of the number of working days.

4. **Figure 4.**

Graph depicts simulated and actual total radiological image data as a function of the number of patients.

5. **Figure 5.**

The distribution of numbers of total 14"×14" X-ray films used for radiological examinations during one working day. The curved line indicates the least square fit of these frequencies to the Gauss function.

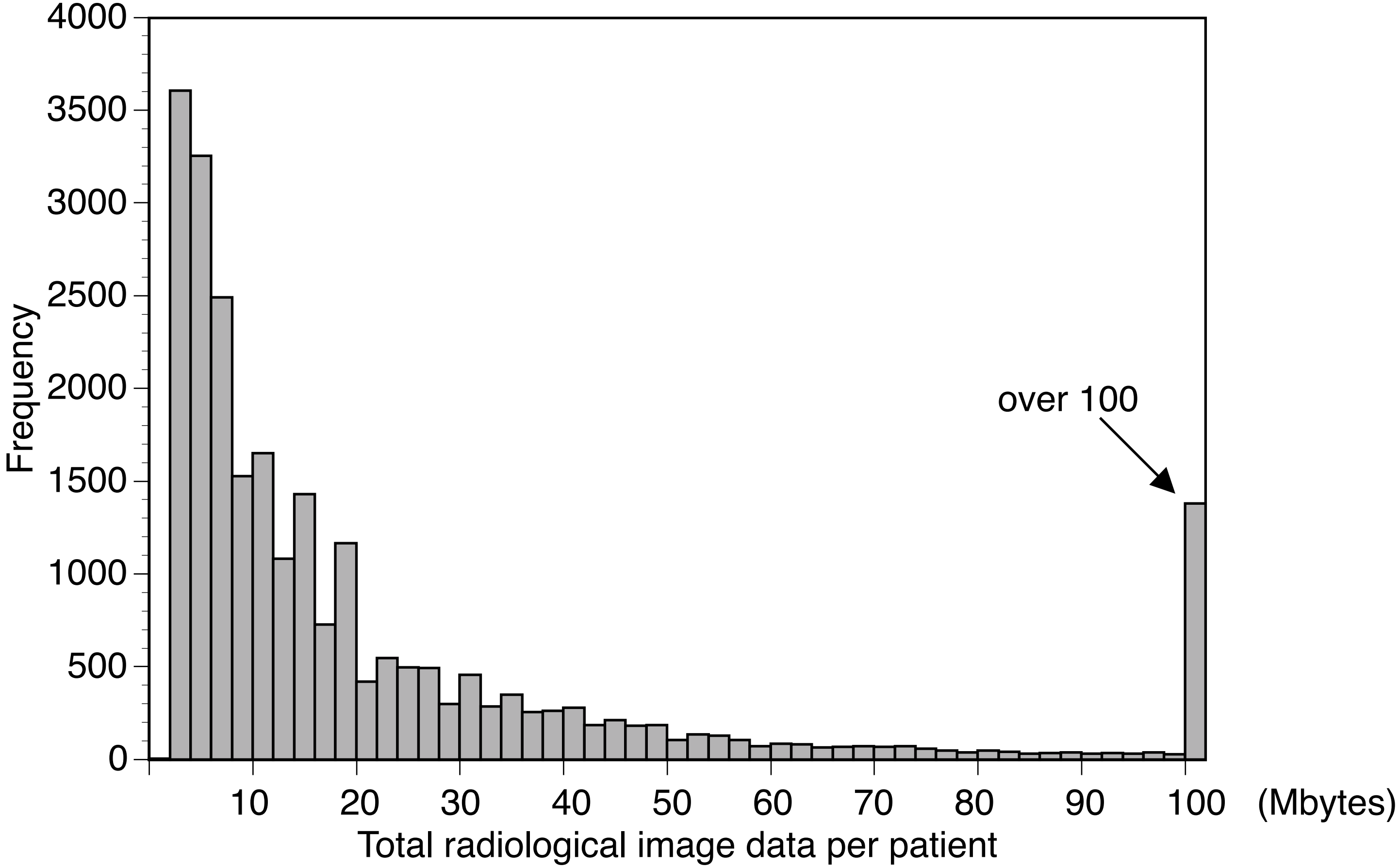
4000
3500
3000
2500
2000
1500
1000
500
0
Frequency
over 100
10
20
30
40
50
60
70
80
90
100
(Mbytes)
Total radiological image data per patient

Figure 1

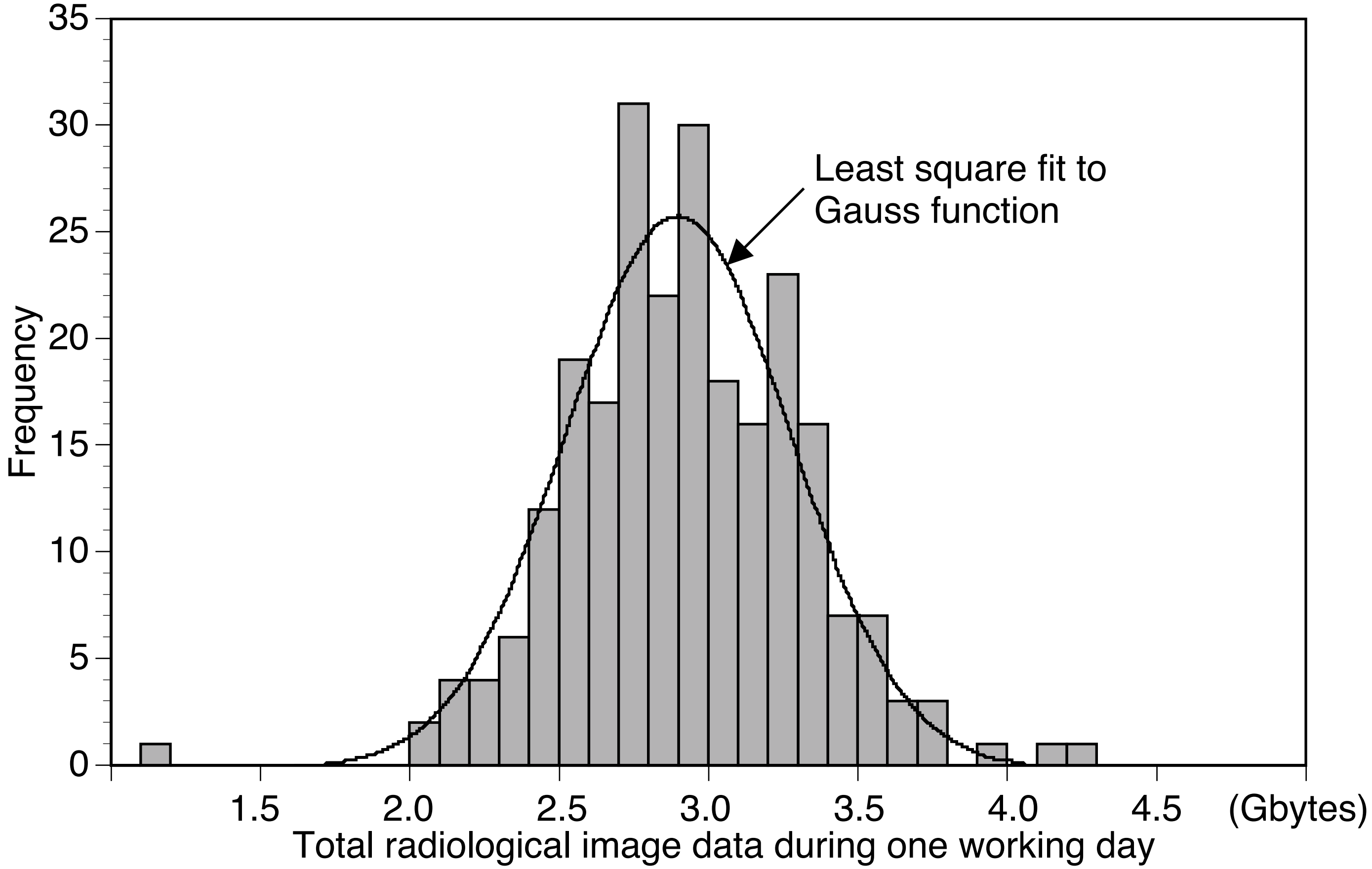
35
30
25
20
15
10
5
0
Frequency
Least square fit to
Gauss function
1.5
2.0
2.5
3.0
3.5
4.0
4.5
(Gbytes)
Total radiological image data during one working day

Figure 2

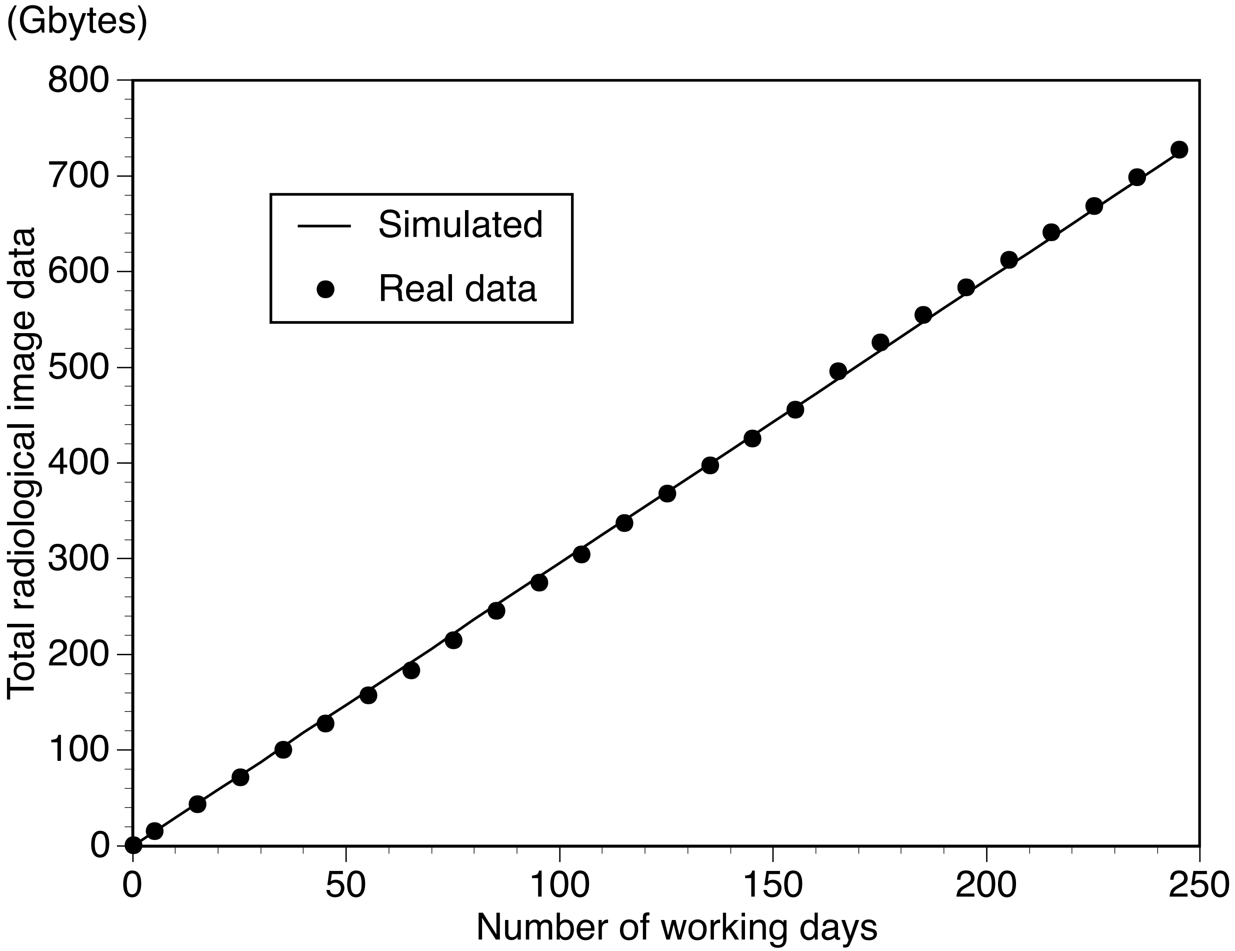
(Gbytes)
Total radiological image data
800
700
600
500
400
300
200
100
0
0
50
100
150
200
250
Number of working days
Simulated
●
Real data

Figure 3

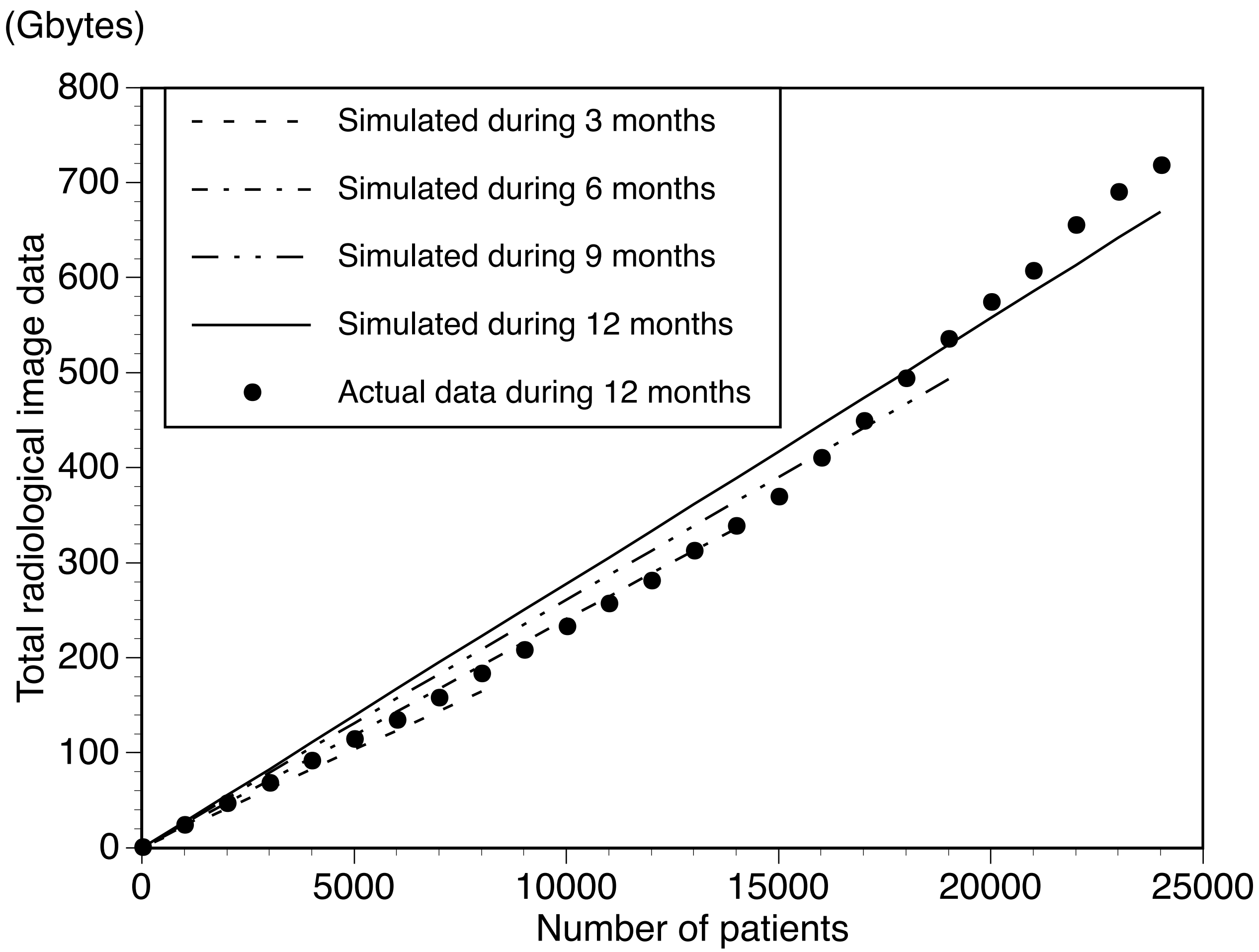
(Gbytes)
Total radiological image data
0
100
200
300
400
500
600
700
800
0
5000
10000
15000
20000
25000
Number of patients
Simulated during 3 months
Simulated during 6 months
Simulated during 9 months
Simulated during 12 months
Actual data during 12 months

Figure 4

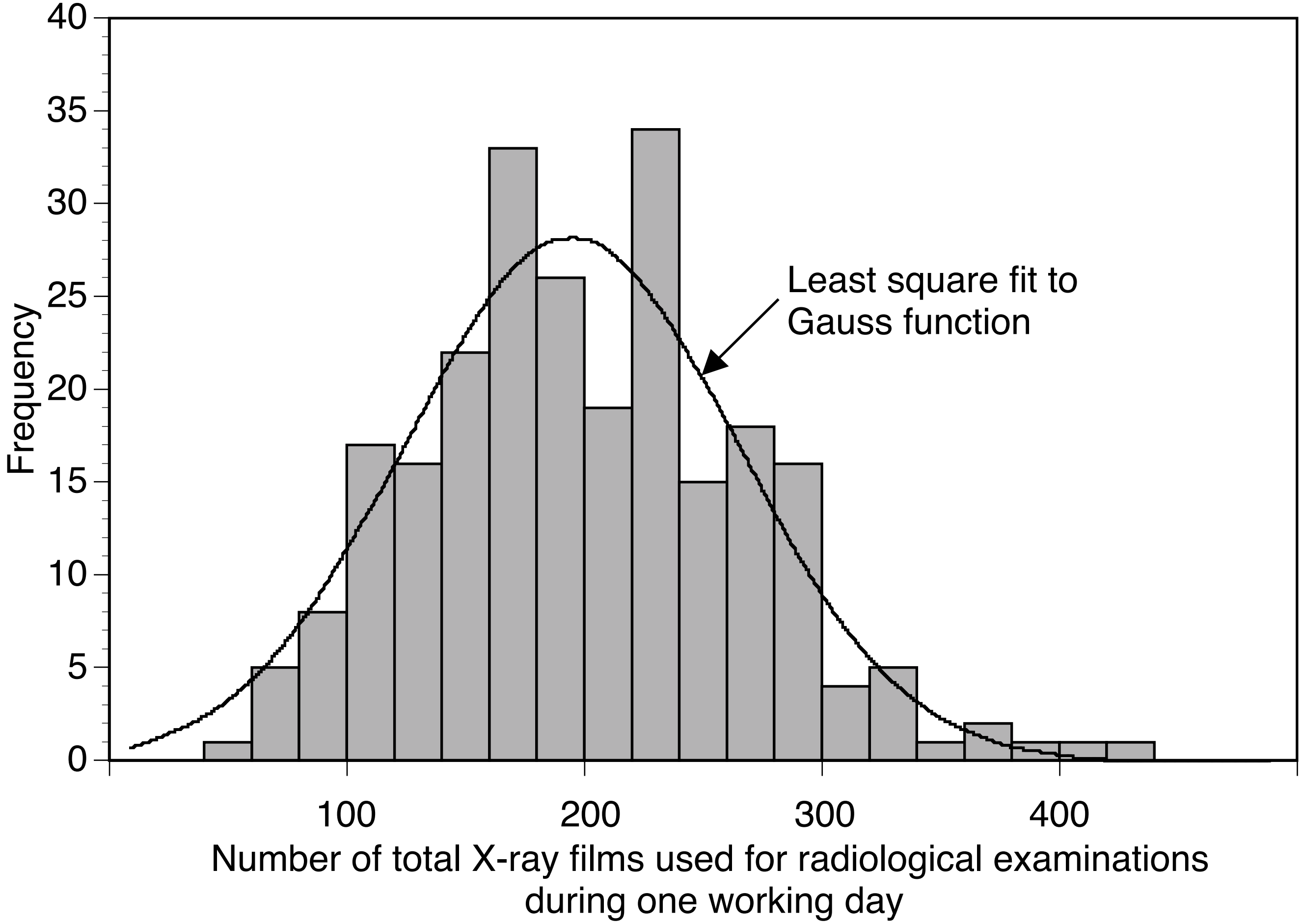
0
5
10
15
20
25
30
35
40
Frequency
100
200
300
400
Number of total X-ray films used for radiological examinations during one working day
Least square fit to Gauss function

Figure 5

Summary

To estimate the size of the media necessary to construct a medical image database system, a statistical analysis was made of the radiological data in our hospital information system (HIS) database; the period of this analysis covered the 12 months from October 1, 1993, to September 30, 1994. During that period, 24,868 patients underwent radiological examinations.

In this present study, we calculated the radiological image data storage required if all radiological images are stored on optical disks handled by the image database. Nuclear medicine examinations, ultrasound, and magnetic resonance images were excluded from this calculation.

The distribution of total radiological (digital) image data storage required per patient over 12 months is not a normal distribution; the mean of these was 30.32 ± 97.44 Mbytes. On the other hand, the distribution of total radiological (digital) image data storage required during one working day is similar to a normal distribution; the mean of these was 2.97 ± 0.78 Gbytes. Therefore, we made the least square fit of these frequencies (as a function of total radiological image data storage required during one working day T) to the Gauss function, $a \exp\left[-\frac{(T-m)^2}{2\sigma^2} \right]$, where a, m, and σ^2 are parameters for which the following estimated values were obtained: $a = 25.9 \pm 1.2$, $m = 2.91 \pm 0.02$, $\sigma^2 = 0.141 \pm 0.015$ (the correlation coefficient for these parameters R^2 is 0.928).

Therefore, the mean amount required for data storage of radiological images during one working day is very useful in estimating the amount required during a given number of working days. In fact, the estimated total radiological image data storage required during a given number of working days was almost equal to the actual amount.

From our investigations, an estimate of the size of the media required for storage of radiological image data during a given number of working days (D days) from an academic 922-bed hospital is given by $2.955 \times D$ Gbytes. Furthermore, an estimate of the size of the media required for D days data is generally given by $T_m \times D$ (where, T_m is the mean amount required for storage of such data per working day). The amount required for storage of medical information other than image data was negligible. Therefore, the size of the media necessary to construct the global image database combined with clinical information is also given approximately by $T_m \times D$.